En la medida en que esta obra suya se ha convertido en un punto de referencia ineludible del filosofar actual, podemos considerar a Derek Parfit como un auténtico clásico viviente. En efecto, su planteamiento del problema de la identidad personal a través del tiempo, en una línea que inauguró Hume en la época moderna, pero que puede vincularse sin excesivos problemas, si queremos adoptar una perspectiva transcultural, con lo más incisivo del pensamiento budista, ha venido constituyendo desde hace ya más de veinte años el centro del apasionado debate intelectual que viene desplegándose en torno a esta cuestión tan importante. Cuestión que enlaza directamente con las problemáticas de la racionalidad práctica y de la ética, y cuyo tratamiento filosófico no dejaría de tener significativas repercusiones psicológicas en todos nosotros, por ejemplo, en nuestra actitud ante el envejecimiento y la muerte. En esta edición española de *Razones y personas* se viene a incluir, a guisa de «epílogo», una definitiva actualización del reduccionismo parfitiano de la identidad personal, redactada más de diez años después. En suma, nos hallaríamos ante una de las obras culminantes del pensamiento de finales del siglo XX: su brillantez y espectacular vigor intelectual no han podido dejar de ser reconocidos ni siquiera por los adversarios más decididos de las sorprendentes tesis que en ella se defienden.

ISBN 84-7774-770-9



RAZONES Y PERSONAS



Razones y personas

Derek Parfit



Razones y personas

www.machadolibros.com

TEORÍA Y CRÍTICA

Colección dirigida y diseñada por
Luis Arenas y Ángeles J. Perona

C/ LABRADORES, S/N. P. I. PRADO DEL ESPINO
28660 BOADILLA DEL MONTE (MADRID)
editorial@machadolibros.com

FOTOCOMPOSICIÓN:
VISOR FOTOCOMPOSICIÓN, S. L.

IMPRESIÓN:
GRÁFICAS RÓGAR, S. A.
NAVALCARNERO (MADRID)

ISBN: 84-7774-770-9
DEPÓSITO LEGAL: M-47.790-2004

DEREK PARFIT

Razones y personas

Traducción y estudio introductorio de
Mariano RODRÍGUEZ GONZÁLEZ



MÍNIMO TRÁNSITO
A. MACHADO LIBROS

ÍNDICE

ESTUDIO INTRODUCTORIO

RAZONES Y PERSONAS

PRIMERA PARTE

TEORÍAS CONTRAPRODUCENTES

Segunda parte
RACIONALIDAD Y TIEMPO

TERCERA PARTE

LA IDENTIDAD PERSONAL

Cuarta parte
LAS GENERACIONES FUTURAS

APÉNDICES

ESTUDIO INTRODUCTORIO

PARFIT O LA VIDA SECRETA DE LAS TEORÍAS

Mariano RODRÍGUEZ GONZÁLEZ
Universidad Complutense de Madrid

Uno de los más notorios supuestos de la obra cuya edición española presentamos es el de que los seres humanos no obran sólo a golpes de capricho o al azar, o por la fuerza ciega del mecanismo, sino que en buena medida tienen la posibilidad de levantar la cabeza por encima del nivel meramente evolucionista de la lucha por la vida, para hacer lo que las teorías que hacen suyas les dicen que hagan, una vez que les han proporcionado determinados fines. Se trata de ser racionales y de ser morales, considerados estos modos de ser como fines formales de las diferentes teorías de la racionalidad y la moralidad. Por eso interesa sobre todo investigar la estructura de estos racimos de teorías, y las relaciones que entre sí mantienen, pues las insuficiencias y las inconsistencias redundarían, a no dudarlo, en nuestro fracaso práctico, racional o moral. Hoy en día, cuando no nos dejamos de alarmar ante la constatación de que, por lo menos aparentemente, la gente se mueve cada vez menos por argumentos, las personas parfitianas se hallarían embarcadas en la apasionante tarea de poner en forma sus razones para salvar de alguna manera sus vidas. Algo que a todos nos debería servir de motivo de esperanza.

La cuestión por la que Derek Parfit ha llegado a ser universalmente reconocido como un clásico viviente, la de la identidad de las

personas a través del tiempo, digámoslo así para resumir, no se plantea desde luego en el vacío: semejante clase de cuestiones *jamás* se podría plantear en el vacío de la reflexión seca y desvinculadamente académica. No estamos ante un problema para el estilete ocioso del filósofo especialista, que en eso de ser especialista contradice el sentido que le puede corresponder a la filosofía, y, por tanto, compromete su imperiosa necesidad para la cultura en un tiempo de especialismos exacerbados. Lo de ir en contra del sentir mayoritario en lo que respecta al problema de la constitución de las personas que seríamos, se inserta en las necesidades concretas de una tradición ética muy particular, la utilitarista que domina en los países de habla inglesa. Se trata de resolver el problema de la racionalidad de la elección moral, en dos palabras, de responder a la crítica tan habitual que se hace al Utilitarismo según la cual éste exigiría un sacrificio excesivo al individuo, que se destina a ser una especie de santo ajeno al mundo del egoísmo racional. Esta insoportable tensión entre altruismo y egoísmo, entre moralidad y prudencia, entre la elección que promueve los intereses del otro y la que nos dicta la teoría del Propio Interés, Parfit la va a intentar aliviar en primera instancia, naturalmente, en el plano propiamente ético, que es el que le corresponde. Tenemos así las dos primeras partes de este libro. Pero luego despuntará la posibilidad de que no se pueda zanjar en ese plano en que se plantea, si antes no se ha decidido la cuestión de qué tipo de entidades son las personas. Y así se nos presenta la tercera parte, que al final vuelve a retomar el problema ético, como exigiría la lógica de la investigación parfitiana.

En (Parfit, 1979), por ejemplo, tenemos una muestra de cómo plantea nuestro autor ese problema del Utilitarismo en el plano propiamente práctico, sin involucrarse en la cuestión metafísica de la condición de persona. Se trata de sondear la posibilidad de que, como ocurre en muchas situaciones vitales cuya estructura corresponde a la que se plantea en los Dilemas del Prisionero, la tesis del egoísmo racional sea contraproducente, esto es, que si cada quien hace lo que es mejor para él, esto sea peor para todos. La elección altruista o abnegada resuelve el problema práctico que conlleva el dilema. Pero no lo elimina, porque subsiste un problema teórico.

Queda en pie el hecho de que la moralidad ha entrado en conflicto con la racionalidad: sigue resultando mejor para *cada cual*, aunque no para *nosotros*, hacer E, la elección egoísta, de manera que para alcanzar las soluciones morales todos tendríamos que actuar irracionalmente. Demostrar, como quieren muchos autores, que, a este nivel teórico de la contraposición cada uno/nosotros, la elección altruista es la racional, no resulta desde luego nada fácil. Parfit nos intenta convencer en este trabajo suyo de que la salida moralista según la cual la prudencia se anula a sí misma, es decir, que incluso en términos prudenciales es la moralidad la que vence, constituiría una opción precipitada. Todo lo más que podríamos decir es que la elección egoísta resulta colectivamente contraproducente.

También es verdad que la prudencia o teoría del Propio Interés, como teoría de la racionalidad que es, tiene sus problemas con otra teoría rival, la instrumental o del fin Presente. Si en relación con la moralidad se enredaba en dilemas interpersonales, con la teoría instrumental se enmaraña en dilemas intertemporales. De lo que se trata no es sino de hacerle la vida incómoda, o hasta imposible, a la teoría del Propio Interés, cerrándole todas las escapatorias en su enfrentamiento con la moralidad y con la teoría del fin Presente. Pero parece que siempre, o casi siempre, se nos acaba escabullendo por entre los espacios que deja libres la red que le hemos echado encima. Hasta aquí, nuestro autor ha tenido que registrar los pasos básicos de la trayectoria vital de ciertas teorías fundamentales sobre la moral y la racionalidad, así como radiografiar sus relaciones conflictivas, y tomar constancia de las refriegas entre las mismas. Pero la fuerza de estos enfrentamientos teóricos —que es el vigor desatado de lo que nos dice qué tenemos que hacer para ser racionales o para ser morales— nos lleva a sospechar que el problema radical, el de las limitaciones que la racionalidad le tendría que imponer a una doctrina moral utilitarista que no atiende a los límites entre las diferentes personas, no es del todo resoluble en el terreno práctico en el que está originalmente planteado. De ahí el desplazamiento al terreno metafísico en el que se representa el drama de la constitución de las personas y de su identidad a través del tiempo. El tratamiento de la cuestión *qué hacer*, desentrañamiento de complejas teo-

rías que luchan unas con otras, nos ha hecho estrellarnos con dos cuestiones metafísicas mayores, la del tiempo y la de la persona (los dilemas que se planteaban en torno a los conflictos entre PI, P y M eran intertemporales e interpersonales): nos vemos llevados por la dinámica interna de todo el asunto al problema de la identidad personal a través del tiempo.

Queda constancia de la posición defendida respecto de las cuestiones de la *naturaleza* y la *importancia* de la identidad personal ya en una serie de artículos de la década de los setenta (Parfit, 1971/1983, 1971, 1973, 1976/1985). Sus tesis se exponen mayormente por la vía del ataque a las creencias de lo que podríamos llamar sentido común. El de Parfit es un auténtico proceso de ilustración que nos llevaría a dar el paso del No Reducionismo al Reduccionismo, y este último se defiende en la forma de una crítica, en verdad demoledora, de aquel. Nuestras intuiciones básicas al respecto serán sistemáticamente desafiadas y minadas por el habilísimo empleo de «experimentos de pensamiento», como por ejemplo el caso de la división, de Wiggins, en el que cada hemisferio del mismo cerebro se lleva a la caja craneal de un cuerpo humano diferente. No podemos sino recordar aquí además otro caso que va a hacer justamente célebre a nuestro autor: el impresionante del teletransporte que se expone al comienzo de la tercera parte de este libro que presentamos. Casos que sugieren que yo sobrevivo como dos personas distintas sin que ocurra que yo soy las dos personas, ni una de ellas y no la otra, pero tampoco está del todo claro que ninguna de las dos. Casos que parecen demostrar que la pregunta acerca de la identidad a través del tiempo no tiene por qué tener siempre una respuesta, y que ponen en obra la ruptura entre *lo que importa* y la identidad personal. Lo que importa es una relación que generalmente está entrañada y oculta por la de identidad —en verdad, para el Reduccionismo ésta *se reduciría* a aquélla— la de la continuidad psicológica, que a su vez consiste parcialmente en *conexividad*, la que, a diferencia de la de identidad, es una relación que admite grados. La palabra «yo» puede emplearse, se nos sugiere, cuando se da el mayor grado de conexividad psicológica. Pero cuan-

do se han reducido las conexiones estaría perfectamente justificado expresarnos diciendo: «No fui yo quien hizo esto, sino un yo anterior», como sugería el gran Proust tan citado por nuestro autor, pudiendo pasar entonces a describir cómo y hasta qué punto estamos relacionados con ese yo anterior. Y pudiera ser que ese yo anterior me parezca ahora un desconocido, que todo lo que ese yo anterior deseaba, creía y admiraba, en suma, cómo vivía y cómo trataba de vivir, hubiera cambiado radicalmente. Lo que Parfit subraya es que no es verdadera la creencia de que haya una persona subyacente que ambos seamos, e insiste además, contra Lewis, por ejemplo, en que estas ideas reduccionistas suyas contradicen frontalmente el sentido común, por cuanto éste considera que lo que importa es la identidad personal. La relación R (continuidad y conexividad psicológicas) es lo que de verdad importa, pero no *es* la misma relación que la de la identidad personal: ésta es todo-o-nada, aquella cuestión de grado, como los experimentos de pensamiento demuestran (aunque el Reduccionismo sostiene que la identidad personal no viene a consistir más que en la relación R, esto sería un asunto diferente). Por tanto, la identidad personal no es lo que importa en la supervivencia.

Para aclarar lo que quiere decir con su Reduccionismo, con lo que alguna vez denominaba la Concepción Compleja, frente a la Simple del No Reduccionismo, Parfit hará un uso continuo de la analogía humeana de las naciones. Al hablar de la identidad de Inglaterra a través del tiempo podemos estarnos refiriendo a la identidad en el sentido lógico, que es del tipo todo-o-nada, como lo es toda relación lógica de identidad, o bien, más realistamente, a la identidad de esa nación en el sentido de su verdadera naturaleza, que, al contrario, sería una cuestión de grado. (Aunque por nuestra parte podríamos pensar que las preguntas «¿Es la Inglaterra Medieval la misma Inglaterra que la actual?», «¿Era también Inglaterra?» carecen propiamente de sentido, o bien se prestan antes que nada a respuestas de orden puramente convencional. «¿Son la Rusia Imperial y la Unión Soviética naciones diferentes o la misma nación?». Sin duda, en parte sí y en parte no, serían las dos cosas.) Pues bien, si nos desplazamos a la Concepción Compleja a partir de

la Concepción Simple que todos al parecer ocupamos naturalmente, la pregunta por la identidad de las personas habrá que responderla como la pregunta por la identidad de las naciones. La identidad de una persona a través del tiempo es sólo *en su lógica* del tipo todo-o-nada, mientras que *en su naturaleza*, al contrario, es cuestión de grado. ¿Soy yo a mis cincuenta años la misma persona que la que alguien fotografió a los diez años, esa foto amarillenta que se conserva en el álbum familiar? En gran parte no, en cierta pequeña parte sí. Somos supervivientes parciales de los adolescentes que fuimos, por decirlo de algún modo llano. Una vez más: «La propuesta es que la vida de una persona puede ser dividida en las vidas de yoes sucesivos. Esto puede hacerse donde se da un marcado cambio en el carácter o alguna otra disminución en la conexividad psicológica. Dónde se haga, queda a la elección del hablante. Se hace con observaciones como esta: "Ese era sólo mi yo pasado"» (Parfit, 1971: 686). Pero distinguir entre el joven del que se enamoró hace muchos años y su cínico marido de hoy, en el caso de la mujer que hizo la promesa al primero de no hacerle caso al segundo cuando le dijera que la rompiese, o entre la Rusia Imperial y la Unión Soviética, por referirnos a dos ejemplos de nuestro autor, no hay duda de que se ajustaría más «a los hechos» que no hacerlo, por mucho que pueda parecer que es algo que se deja al arbitrio del hablante.

El Reduccionismo defiende, en resumidas cuentas, que «el hecho» de ser una persona, como algo distinto a ser un simple animal, *consiste en* tener otras propiedades más específicas, como por ejemplo, tradicionalmente, la racionalidad, propiedades que se tienen en diferentes grados. Por eso a esta posición se la puede llamar «compleja». Mientras que la concepción que se opone al Reduccionismo simplemente sostiene que ser una persona no consiste en nada distinto de ser una persona, que la condición de persona es un «hecho adicional profundo», aparte de la mera continuidad psicológica con base cerebral, que no puede darse en diferentes grados (Parfit, 1973: 137). Es por tanto la concepción «simple». Exactamente igual ocurriría con «el hecho» de la identidad personal a través del tiempo —no se puede separar el problema del criterio

de identidad personal del problema de las condiciones de la *personhood* o cualidad de persona—. Para el Reduccionismo, cuyo representante quizás más señalado sea Parfit, este hecho *consistiría en* la continuidad física y psicológica, es decir, en otros hechos más específicos que desde luego son en parte cuestión de grado. Para la posición contraria a la reduccionista, en cambio, el hecho de la identidad personal a través del tiempo tiene una naturaleza *especial*, de algún modo *profunda*, en el preciso sentido de que sería totalmente independiente de los otros hechos más específicos: se añadiría a ellos, por así decir, de manera que o bien se da por completo o no se da en absoluto (Parfit, 1973: 138). De forma similar, lo que es importante en la identidad personal serían las dos relaciones mencionadas de la continuidad y la conexividad. Son lo que interesa en la supervivencia. La lógica de la continuidad es también todo-o-nada, pero la conexividad, que la continuidad implica, es evidente que admite grados, de manera que, como diría el propio Parfit, *en su naturaleza* la misma identidad personal tiene que admitir grados. Con lo que vamos a insistir una vez más en la doctrina contraria a la del sentido común. En todo caso, parecerá que si aceptamos la Concepción Compleja la identidad personal importa menos porque implica menos.

El verdadero problema radica en comprender cómo es posible que, como él mismo afirma una y otra vez, con su posición radicalmente reduccionista Parfit no pretenda negar la *realidad* de las personas (¿tal vez no se atreve a declararlo explícitamente porque al fin y al cabo equivaldría a negar *nuestra* realidad?): a su juicio no seríamos, en sentido estricto, series de sucesos, simples cadenas de pensamientos y de acciones, sino más bien pensadores y agentes. Ahora bien, lo que el pensador oxoniense añade, el sentido de la tremenda matización que hace seguir a esta concesión, es lo que resultaría ciertamente problemático, al menos a las alturas del año en que la realiza: «Pero nosotros consideramos esto un hecho gramatical» (Parfit, 1973: 158). Somos distintos de nuestros cuerpos, de nuestras acciones y nuestras experiencias, ¡pero sólo en un sentido *conceptual*! Sin duda que es con este problema contra lo que tendrá que luchar en lo sucesivo, y hasta por lo menos el año en que se publi-

ca el escrito que, por una gentileza de su propio autor verdaderamente digna de agradecer, incluimos al final, como epílogo, en esta edición de su obra capital (Parfit, 1995b). Pero en un principio la solución fue poner el problema en la cuenta de nuestra incapacidad natural de asumir la solución reduccionista, porque Parfit tenía muy claro que la raíz de esta, el rechazo del *deep further fact*, tenía que implicar la afirmación de que somos distintos de nuestras experiencias sólo porque así lo quiere la Gramática.

El paso crucial de la Posición Simple, en la que estaríamos instalados en tanto que somos en cierto sentido siempre *ordinary folk*, a la Concepción Compleja reconocida como verdadera en el proceso de ilustración filosófica, tiene consecuencias de amplísimo alcance, en primer lugar, para nuestras teorías de la racionalidad y de la moral. Nada más y nada menos: para nuestra misma visión de la vida práctica de todos los días, en sus dimensiones tanto prudencial como ética. Ya comentamos que, para toda la notable tradición de los Bentham, J. S. Mill y Sidgwick, había venido constituyendo un problema de decisiva importancia el de la justificación racional de la conducta abnegada. Maximizar la felicidad general puede enfrentarnos en ocasiones a la conveniencia o incluso la necesidad del autosacrificio, y lo indudable para casi todos estos pensadores es que la racionalidad práctica exige en principio seguir la orientación del propio interés, interprétese este como se interprete. Toda desviación del propio interés levantaría *ipso facto* la sospecha de haber incurrido en crasa irracionalidad. A la pregunta del individuo concreto: ¿y por qué iba a tener que renunciar a mi propio placer en aras de la felicidad general?, no ha resultado nada fácil encontrarle una respuesta medianamente satisfactoria en el ámbito de una filosofía moral que le tiene alergia al misticismo y que se inclina decididamente por los «hechos» y el naturalismo como es la utilitarista, y esto constituye un colosal obstáculo desde el momento en que es esta misma concepción ética la que hace de la mayor felicidad del mayor número el principio supremo de la acción humana (Scarre, 1996).

Nada más exponer por vez primera su teoría de la identidad personal a través del tiempo, Parfit no pudo por menos que declarar

solemnemente que «el principio del interés propio no tiene ninguna fuerza» (Parfit, 1971/1983: 34). Luego el tono de contundencia desaparece. El cambio de creencias respecto de la identidad de las personas tendrá meramente el efecto de *debilitar* nuestra creencia natural en la teoría del Propio Interés: sobre *quién* recaiga el beneficio o la pérdida, la felicidad o la desgracia, no tiene al fin y al cabo tanta importancia como la cualidad y la intensidad de la experiencia misma, vistas las cosas desde una perspectiva que ha dejado de tomarse en serio, como ilusoria y falsa, la idea del hecho adicional profundo. Si no hay yo posesivo cartesiano no es tan decisivo quién saldrá ganando, sino la ganancia misma, en cualquier caso. Si seguimos firmes en la falsa idea de la Concepción Simple, por el contrario, *la cuestión del quién* conservará su importancia, y el Utilitarismo continuará enfrentado a la dificultad de siempre. Mientras que desde la Concepción Compleja resultará más plausible presentar los «datos morales» en una forma impersonal. El incremento en plausibilidad y apoyo no quiere decir, desde luego, que la nueva concepción *implique* la tesis impersonal utilitarista. No hay nada en una nación diferente de sus ciudadanos y no hay nada en una persona distinto de sus experiencias: de manera que cuando descubrimos esto deja de ser tan importante la nacionalidad de una persona, como deja de ser especialmente significativo quién es el poseedor de estas experiencias (Parfit, 1973: 158). Al final se viene a representar una escena mucho más modesta, casi por vía de sugerencia: Rawls le habría reprochado al Utilitarismo no tomarse en serio la distinción entre personas, y, si en efecto esta distinción es el hecho básico de la filosofía moral, estaríamos ante una objeción de muy grueso calibre. Ahora bien, habría una concepción de la naturaleza de las personas, la parfitiana, que serviría para proporcionar alguna defensa al Utilitarismo (no una defensa *suficiente*, reconoce con humildad nuestro autor).

Puede que no sea tan racional como habíamos venido creyendo, en suma, orientar nuestra conducta en el sentido de favorecer más y mejor nuestros propios intereses. Sobre todo, puede que no haya por qué denunciar la irracionalidad de toda actitud diferente de la del egoísmo «racional». Es este egoísmo el que nos hace ciegos a las

preocupaciones del otro, haciendo nuestras vidas estrechas y mezquinas, encerrándonos en ese túnel de cristal del que confiesa estar harto Parfit, al final del cual sólo habría noche y nada. Por eso no parece importarle mucho a nuestro autor la objeción de que si nos preocupamos menos por nuestra propia identidad nos preocuparemos menos también por la identidad del otro (Parfit, 1986: 837).

Tornar menos creíble la teoría del Propio Interés, disminuir su potencia cultural milenaria declarando la no existencia de egos sustanciales, el fin del error, tendría desde luego también efectos en nuestras emociones y actitudes, determinaría cambios psicológicos que, juzgando sobre todo por su propio caso, Parfit considera beneficiosos (¡y pensar que no encontrar al yo casi pone al borde de la desesperación a Hume!: parece que por lo menos en este terreno de la identidad mucho va en sensibilidades). Tenemos en primer lugar nuestra actitud ante el futuro, la sempiterna y trabajosísima *cura sui*, la vigilancia siempre alerta por nuestra suerte personal. En las páginas de Parfit afloraría la voz de todo este milenario cansancio, con el eco del Sermón de la Montaña —sermón que contrasta con la acusación, de raigambre nietzscheana, de que el Cristianismo representa históricamente la consagración definitiva del cálculo personal y del propio interés, en su multiplicación al infinito apoteósico de la eternidad— el canto a la despreocupación, al dejarse llevar, al desposeimiento, al abandono de sí, el budismo de la extinción del cuidado del yo. Una nueva vida, sin duda, de la que no se hace mucha propaganda explícita, sólo nos son sugeridos sus encantos.

Todas las partes temporales de nuestra existencia no son *nuestras* por igual, la conexividad habíamos visto que es cuestión de grado, o, para decirlo a la inversa, todas las partes de nuestra vida son nuestras por igual, pero esto es sólo una verdad trivial, gramatical, que no supone nada profundo, no importaría nada en realidad (Parfit, 1971: 686). Muerde menos, de este modo, el remordimiento. Y cosas tan aparentemente naturales como la tristeza de envejecer y el miedo a morir le parece a nuestro autor que pierden buena parte de su sustancia cuando atacamos a fondo, y con éxito, la Concepción Simple de la identidad personal a través del tiempo, porque en gran medida se basan en esa teoría. Psicológicamente

hablando, para Parfit no cabe duda de que son nuestras creencias las que en buena parte determinan el aspecto de nuestras emociones. Por eso la filosofía tiene un alcance psicológico y humano tan grande. Por eso no es cosa de niños andar criticando unas teorías y defendiendo otras... hay que tener cuidado de lo que se dice, sin duda.

Uno puede pensar que no es un objetivo menor de todo el pensamiento parfitiano liberarnos del miedo a morir. No es lugar este para tratar tema tan tremendo, y habría que matizar muchísimo, como hace el mismo autor, que en este terreno como en ningún otro no se hace ilusiones. Sometida al tratamiento al que la somete, hay un momento en que la muerte como tal parece desvanecerse. No que sea impensable, sino que al pensarla resulta casi nada. «Después de que haya pasado un cierto tiempo, ninguna de las experiencias que ocurrirán estará conectada, de modos determinados, con estas experiencias presentes»: en este hecho y en nada más consistiría haber muerto (Parfit, 1986: 837). Con lo que temer a la muerte, temerla directamente a ella, a ella en sí misma, no sería algo muy apropiado para el humano que ha decidido guiarse por la razón. Pero baste decir que, en resumidas cuentas, se recomienda el Reduccionismo no sólo, aunque sí fundamentalmente (Parfit al fin y al cabo no es un psicólogo), porque se piense que es verdadero, sino también porque se está convencido de que es mejor en general para nosotros que sea verdadero. No se trata entonces de que se cumpla la verdad y perezcamos, como quiere el lema de los suicidas por el conocimiento, sino de que nos viene muy bien que la verdad sea la que es, y por tanto conocerla. Habríamos descubierto que la relación en que me encuentro hoy conmigo mismo mañana es igual de «íntima», no en absoluto más, que la relación en la que me encontraría con una reproducción mía obtenida por clonación, por ejemplo.

Hay otras repercusiones más técnicas y tal vez de menor interés directamente humano, pero en absoluto desdeñables, todas en el sentido general de una confirmación de las ideas utilitaristas. Se refieren sobre todo al merecimiento y al compromiso, a las promesas y la justicia distributiva. Vamos a mencionarlas nada más. Está

claro que merecer premio o castigo parece presuponer la identidad personal, o la importancia moral de la identidad personal. Y que sin duda la Concepción Compleja haría menos plausible tal presupuesto. Así que en el Reduccionismo está contenida la tendencia a debilitar los principios del merecimiento. Pero también, por el mismo razonamiento, el principio que fundamenta la obligación de respetar los compromisos y de cumplir las promesas que hicimos en el pasado distante. No hay por qué pensar, insiste Parfit en ello, que su concepción de la identidad a través del tiempo termine con el respeto a las promesas y a los compromisos, y liquide toda justificación para aplicar castigos y otorgar recompensas, así como para encontrarle un sentido al concepto de culpa. Según él, la continuidad psicológica bastaría para que todos estos conceptos morales retuvieran su importancia, si bien debilitada. Lo que a nuestro autor parece importarle especialmente es el hecho, para él indudable, de que todas estas repercusiones irían directamente a favor del Utilitarismo. Ya no sería, sin embargo, tan directo el sentido pro-utilitarista de las consecuencias del cambio a la Concepción Compleja en el caso de la justicia distributiva y de la compensación. Lo que buscan los utilitaristas es «maximizar» —la mayor suma neta de beneficios menos cargas, cualquiera que sea su distribución— de manera que rechazan los principios distributivos, desatendiendo los límites entre las diferentes vidas. Hay dos clases de distribución, entre las vidas y dentro de las vidas. Y hay dos maneras de abandonar los principios distributivos, o bien no darles ningún alcance o bien no darles ningún peso. La Concepción Compleja justificaría tal vez conceder un alcance más amplio a estos principios, extendiéndolos al interior de cada vida (de la misma manera que pensamos que un sacrificio de una persona no sería compensado por un beneficio de otra, seríamos del parecer de que un sacrificio impuesto a un menor no podría ser compensado por un beneficio obtenido por su yo adulto). Pero sin duda apoyaría más enérgicamente el debilitamiento de los mismos (la compensación presupone la identidad personal, así que desde la Concepción Compleja podemos pensar que el hecho de la compensación es en sí mismo menos importante moralmente). De forma que el efecto

neto del cambio a esta concepción también iría en la dirección de un fortalecimiento del Utilitarismo (Parfit, 1973: 153). Con todo, buen cuidado se pone aquí, sin duda en parte por la influencia de las críticas recibidas, en equilibrar el impacto de estas repercusiones. Por ejemplo, aunque si nos hiciésemos reduccionistas tendríamos que reconocer que no puede haber una compensación absoluta a través del tiempo, eso no quita que pudiéramos pensar que sí que podría darse «cuasi-compensación»: igual que podemos ser cuasi-compensados de nuestros sufrimientos por las ganancias que vayan a recibir los seres que a los que queremos, podemos ser cuasi-compensados de nuestros sufrimientos por las ganancias que obtengamos en otras partes de nuestra vida que ahora nos preocupan. Esto es, aunque seamos reduccionistas seguiríamos estando preocupados por nuestro futuro *en cierta medida*, responde nuestro autor a sus críticos (Parfit, 1986: 862).

En muchos de sus trabajos ha conectado Parfit los asuntos de la identidad personal y de la moralidad, aunque no sea en un sentido tan directo y estricto como en el esquema básico que venimos de analizar. Y en todos ellos se ha mostrado sumamente receptivo a sus numerosos críticos, desde recopilaciones tempranas como las de Perry (1975) y A. O. Rorty (1976), hasta obras colectivas en que se examinaba *Razones y personas* desde todos los puntos de vista imaginables, como el número 96 de *Ethics*, en cuyos «Comments» nuestro autor daba respuesta a numerosas voces críticas, o como la edición tardía de J. Dancy (ed.) (1997) *Reading Parfit*, que en vez de facilitar la lectura de su obra capital introduciendo de manera accesible los diferentes temas abordados en ella, lo que hace más bien es presentar enjundiosas impugnaciones a sus principales tesis, a veces impugnaciones a la totalidad. El talante exhibido ante la recepción crítica de sus teorías lo representa muy bien el siguiente fragmento, en el que, después de resumir en lo esencial las objeciones presentadas en *Ethics* por diversos autores a *Razones y personas*, nuestro autor nos demuestra una vez más que si el filósofo, como decía Heidegger, siempre se encuentra pensando en rigor *lo mismo*, ese pensar lo mismo por su parte tiene el carácter de un *work in progress*

impulsado en su dialéctica interna por las observaciones críticas de la comunidad filosófica de referencia: «No trataré de resumir esta larga conclusión, pero parece que vale la pena enumerar mis conclusiones. Wolf señala que si nos volviéramos reduccionistas, esto tendría ciertos efectos negativos. Kagan advierte una brecha en mi defensa de la Teoría del fin Presente. Gruzalski muestra que, dado que yo no resolví el Problema Sorites, mi discusión de los efectos imperceptibles no respondió a todas las preguntas que hice surgir. Kuflik corrige mi descripción de la Moralidad del Sentido Común. *Hay desde luego muchas otras maneras en que mi libro necesita ser revisado* [cursiva mía]» (Parfit, 1986: 862).

Se entrelazan también moralidad e identidad personal en el planteamiento y la discusión del tema de la justicia intergeneracional. No se puede dejar de constatar en este punto que, aunque el trabajo de referencia sea el artículo seminal de Narveson de 1967, el que llevaba el nombre de «El Utilitarismo y las nuevas generaciones», así como que al que debemos la primera discusión sistemática de nuestras obligaciones para con las personas futuras sea desde luego J. Rawls, haya sido la obra de Parfit, en especial en la forma madura que asume en la cuarta parte de *Razones y personas*, la que ha definido para todos los estudiosos posteriores el sentido en que se plantean los problemas de cómo podemos y debemos relacionarnos con las personas futuras (Meyer, 2003). Esta cuarta parte titulada «Las generaciones futuras» se ha venido considerando con excesiva frecuencia la menos importante del libro, y ha sido en consecuencia la menos estudiada, sin duda, pero frente a este equivocado modo de ver las cosas se ha levantado, sin embargo, la voz de quienes insisten en que «es uno de los trabajos más ricos y más profundos de la filosofía contemporánea» (Temkin, 1997: 290): en ella los argumentos de Parfit se despliegan con una originalidad verdaderamente impactante, enfrentándose directamente al núcleo de nuestras creencias más profundas.

Se pretende aquí, entre otras cosas, discutir el Problema de la No Identidad o las elecciones de diferentes personas: las personas que en la actualidad se hallan viviendo pueden afectar con sus acciones a la existencia misma de las personas futuras, o a su número e

identidad, desde el momento en que adoptamos la concepción genética de la identidad personal, según la cual la identidad de una persona se hallaría, al menos en parte, constituida por el ADN que la persona tiene como consecuencia de qué óvulo fue fertilizado por qué espermatozoide en la creación de esa persona, de manera que nuestras acciones tienen un efecto en la identidad genética de las personas futuras en la medida en que afectan al hecho de a partir de qué pares particulares de células esas personas futuras se originarán. Que yo sea yo, que exista yo, depende de haber sido concebido en el espacio de un mes alrededor del día en que de hecho fui concebido. Algo tan importante aparentemente como mi individualidad única e irrepetible dependería parcialmente de un hecho tan contingente y nimio como el del día concreto de mi concepción. Así, el problema se plantea en los siguientes términos: ciertas elecciones nuestras pueden tener efectos muy negativos en las personas del futuro, efectos que nos aportan razones para no hacerlas; pero puede ocurrir que sea predecible que si no tomamos tales decisiones estas personas futuras particulares nunca vayan a existir, de manera que tomarlas no va a ser peor para ellas. Para Parfit, esto no elimina nuestras razones morales para no hacer esas elecciones. Es su Tesis de la No Diferencia, para la cual las razones que tenemos para no perjudicar a futuras personas posibles, aquellas que podrían ser concebidas, son tan fuertes como las que tenemos para no dañar a personas reales, las ya concebidas. Y ¿cuáles serían estas razones? ¿Cómo podríamos explicarlas? No pueden ser explicadas completamente ni recurriendo a los intereses de la gente ni a sus derechos, sino que necesitamos una nueva teoría que por ahora no conocemos. Asimismo, otro enigma que Parfit intenta descifrar es el de la Asimetría: mientras que los candidatos a ser padres no tienen ninguna obligación de engendrar por consideración a los intereses de los posibles futuros hijos resultantes, sí que tendrían la obligación de no traer al mundo hijos que se pueda pronosticar que vayan a llevar una existencia miserable. De esta intuición de sentido común se derivarían consecuencias paradójicas que nuestro autor se aplicará a examinar pacientemente, con esa casi omnipotente paciencia que es la suya.

La filosofía moral tiene que ver necesariamente con las personas y su identidad porque aquí nos movemos por regla general en una concepción ética que sería *person-affecting*, o sea, aquella para la que la cualidad moral de una acción tiene que ser estimada sobre la base de cómo afecta a los intereses de las personas. Por eso se puede detectar la implicación personalista, si así lo podemos decir, en la mayoría de los trabajos éticos de Parfit, aun en aquellos que aparentemente nada tendrían que ver con la identidad personal. Como el que se abre con la pregunta de cómo podemos hacer la mejor distribución, si tratando de conseguir la igualdad entre las diferentes personas, o antes bien dando prioridad a las que están en la peor situación, un trabajo de argumentaciones cristalinas que se dedica a dibujar las complejas relaciones entre las posiciones utilitaristas, por un lado, y las igualitaristas y las que buscan priorizar los intereses de los menos favorecidos, por otro (Parfit, 1995a).

La influencia de un filósofo contemporáneo se mide sobre todo constatando cómo ha podido erigirse su obra en el centro de importantes debates a los que han contribuido pensadores procedentes incluso de tradiciones filosóficas diferentes, y desde luego la intensidad con la que las críticas han arreciado contra sus posiciones es asimismo un infalible índice de su relevancia y su centralidad. La obra de Derek Parfit es una perfecta ilustración de la utilidad de estos dos indicadores. No hay discusión del problema de la identidad personal a través del tiempo a partir de los años setenta del siglo XX que no la tome como uno de sus referentes básicos, en muchas ocasiones para someterla a todo tipo de refutaciones más o menos supuestas o más o menos logradas. Nuestro autor sabe muy bien que su tesis va en contra de nuestras creencias «naturales» —fomentadas por nuestra cultura milenaria— y que en muchos casos desequilibran la dinámica más habitual de algunas de las emociones más constitutivas de nuestra vida psicológica. Tal vez esto explique toda esa filosófica formación de combate que se ha venido organizando en los últimos treinta años contra las ideas que alcanzaron la más nítida expresión en 1984 en *Razones y personas*. No muchos parfitianos podemos encontrar en las publicaciones internacionales, si queremos decir la verdad, pero los que insisten en contradecir a Parfit son legión, y su número y la calidad de sus objeciones dejan pocas dudas sobre la considerable importancia del filósofo de Oxford, y la relevancia de su revitalización de un problema secular, profundo y repleto de implicaciones éticas, jurídicas y hasta políticas.

Desde la tradición filosófica a la que pertenece su obra, nutridas y de índole diversa son las críticas que la razón analítica ha dirigido contra las tesis de nuestro autor. Pero nos vamos a referir sólo a dos, sin duda de las más sobresalientes, y aparecidas ambas en la antología de J. Dancy, si bien la primera había visto la luz en 1985 en *Mind*, como magistral recensión de *Razones y personas*. Coinciden en rastrear contradicciones, por regla general, pero ya dejó dicho Ricoeur, en la revisión crítica a la que para cerrar estas páginas nos referiremos, que el pensamiento de Parfit, un pensamiento de extraordinario vigor, resulta inatacable si lo abordamos en su propio nivel analítico. De la riqueza del escrito de Shoemaker (1984/1997), extraemos nada más que dos consideraciones fuertemente críticas dirigidas contra la misma idea parfitiana de la identidad personal a través del tiempo, y algunas observaciones que desmentirían las supuestas repercusiones de la misma. La primera consideración se nos antoja inconsistente. Desde una perspectiva que habría que denominar funcionalista, lo que Parfit tiene que considerar que constituye la existencia de una persona es simplemente el sistema de los estados y los procesos mentales. Ahora bien, según Shoemaker, tal consideración implicaría el reconocimiento de una dependencia ontológica necesaria de las experiencias psicológicas respecto de la existencia de *personas* o sujetos mentales. Pero me parece que esta implicación no está nada clara, como tampoco lo estaría, en consecuencia, que la descripción impersonal de las personas a la que se tiende tenga que significar en el fondo la disolución de la mente, su proceder como tal descripción impersonal en términos simplemente físicos o funcionales. ¿No habíamos partido, desde el comienzo, de una funcionalización de la mente? Y no encuentro que la impersonalidad que busca Parfit sea incompatible con la concepción general del Funcionalismo, sino todo lo contrario.

Pero lo que Shoemaker considera más criticable no es esto, sino el sentido de la refutación parfitiana del No Reduccionismo, que procede como si la concepción del ego cartesiano pudiera haber sido verdadera, resultando, sin embargo, falsa, según nos llevaría a creer todo lo que sabemos. Es decir, parece que para Parfit la falsedad del No Reduccionismo es empírica más que *a priori*. Y no se trata tanto, evidentemente, de que Shoemaker esté convencido de que este es incoherente en sí mismo y, por tanto, no podría haber sido verdadero en absoluto. El problema radicaría en que Parfit trabaja muy a menudo la discusión del No Reduccionismo empleando casos imaginarios, lo que produciría la impresión de contradecir su pretensión en relación con él, porque entonces parece que lo que de verdad está en juego, en consecuencia, son una posibilidad y una necesidad conceptuales antes que nomológicas, o metafísicas en el sentido de Kripke. Hay aquí sin duda motivo de crítica, y esto ha sabido verlo muy bien Shoemaker, aunque no podamos olvidar que Parfit también ha pretendido mostrar la posibilidad de la verdad de la tesis del ego cartesiano empleando otros experimentos mentales de no menor significación. Si respetamos el procedimiento argumentativo de los casos imaginarios como procedimiento general, habría que aplicarse a la tarea de mostrar en concreto por qué algunos no funcionan como se espera cuando tratan un asunto del que estamos convencidos que no se presta a este tratamiento porque el autor supone que es un asunto de naturaleza empírica más que conceptual.

Por lo demás, somos de la opinión de que, contrariamente a lo que parece pensar Shoemaker, habría un sentido en que resulta intuitivamente correcto afirmar que si el No Reduccionismo fuera verdadero, la preocupación especial por nuestro futuro estaría más justificada racionalmente que en caso contrario, como también las nociones morales de merecimiento y compromiso tendrían una aplicación menos problemática, y los principios distributivos de justicia un peso mayor. Aparte de eso, si bien es verdad que, como creo que en algún momento reconoce el mismo Parfit, el Reduccionismo puede llegar a hacerse compatible con la Teoría del Propio Interés, en la medida en que la identidad personal consistiría en la famosa

relación R, y ésta no vacía de sentido *completamente* a la preocupación por nuestro futuro, esta compatibilidad resultaría mucho más «natural» en el caso de la posición contraria.

J. McDowell (1997), en su artículo contra la pretensión de someter al tratamiento reduccionista el punto de vista de la primera persona, insiste en que, en relación con lo que él mismo llama «el fenómeno de Locke» (el hecho de que a las personas la *consciousness* les proporcione una perspectiva interna sobre su propia persistencia en el tiempo), tenemos que desenmascarar como totalmente falso el que la alternativa a la que nos enfrentemos sea la de reducción o admisión del ego puro cartesiano. Habría por el contrario una tercera posibilidad, que es la que McDowell defiende y que echa de menos en la obra de Parfit: que exista, de forma continua, una entidad a la que conocemos como *persona*, uno de cuyos aspectos sea precisamente esa continuidad de conciencia. El Reduccionismo al estilo de Parfit sería falso, pero no porque la identidad personal suponga un hecho adicional profundo, sino porque no hay en absoluto ningún sustrato más básico al que se pueda retrotraer por análisis. Hay que suponer entonces que la condición de persona sería ella misma strawsonianamente inanalizable. Pero lo esencial en la crítica que ahora atendemos es la denuncia de un rasgo que por lo visto sería común a Parfit y a los cartesianos, la idea de que la *consciousness* es «autocontenida», la convicción de que el fenómeno de Locke debe ser entendido de forma aislada, con independencia del contexto de hechos que da sentido a la continuación de las vidas humanas en el tiempo. Pero ocurre que sólo ese contexto objetivo es capaz de hacer inteligible que la conciencia continua nos presente una identidad a través del tiempo. Haber prescindido Parfit de él le sirve a McDowell de ocasión para practicar la hermenéutica de la sospecha: impulsaría las ideas de nuestro filósofo la tentación de trascender la finitud de la vida humana individual, nuestra misma realidad de *animales* racionales, y en esta fantasía tan usual no habría que ver en absoluto una supuesta liberación de los prejuicios de la razón práctica, sino todo lo contrario, una distorsión filosófica más, «impuesta a la reflexión sobre la razón y la vida humana por nuestro olvido, como es de entender un olvido buscado por encima

de todo, de cómo mantener una concepción firme e integrada de nosotros mismos como animales racionales» (MacDowell, 1997: 248). Sería el curso que nuestra vida va trazando entre las cosas del mundo lo que constituye nuestra identidad como personas a través del tiempo, como personas terrenales, y esto es precisamente lo que ha querido negar la tradición de la conciencia autocontenida, con la intención de levantar demasiado la cabeza sobre el nivel biológico en el que nos encontramos como especie animal que somos. La crítica de MacDowell sólo parece hacer mella en el discurso parfitiano en la medida en que lo impugna en su totalidad, casi con un pie fuera del ámbito que le es propio. Porque en ese ámbito no se puede pasar por alto que los estados intencionales representan el mundo y significan lo que las causas mundanas determinan: no hay manera de hurtarse al contexto objetivo del que pretendía hacer abstracción toda esa concepción de la conciencia autocontenida, a menos que se trate de una conciencia sin ningún contenido, o sea, propiamente, una no-conciencia.

Desde el terreno que corresponde a la razón trascendental, pasando a otro orden de críticas, se ha podido aseverar que Kant habría refutado a Parfit (S. Blackburn, 1997). Fue Ch. Korsgaard la que, en un trabajo justamente célebre publicado en 1989, llamó nuestra atención sobre la existencia de una clara diferencia entre dos conceptos de *persona* que no podían ser confundidos: las tesis del utilitarista Parfit sólo tendrían sentido en el interior del marco de uno de ellos pero se revelaban totalmente erradas desde la perspectiva del otro: «Una persona es tanto activa como pasiva, tanto un agente como un sujeto de experiencias. Los filósofos morales utilitaristas y kantianos, sin embargo, ponen un énfasis diferente, de una forma que les caracteriza, sobre estos dos aspectos de nuestra naturaleza. El utilitarista subraya el lado pasivo de nuestra naturaleza, nuestra capacidad de estar contentos o satisfechos, y está interesado en lo que nos sucede. El kantiano subraya nuestra capacidad de actuar y está interesado en lo que hacemos. De manera alternativa, podemos decir que el utilitarista se concentra en primer lugar en las personas como objetos de interés moral, y pregunta "¿qué debería hacerse por ellas?", mientras que el kantiano se dirige al agente

moral, que se pregunta "¿qué debo hacer yo"?» (Korsgaard, 1989: 101). Las personas pueden ser contempladas como «*locus* de experiencias», tal y como hace Parfit siguiendo a toda la tradición a la que pertenece, pero también como seres fundamentalmente activos, comprometidos en sus acciones. Lo que tenemos que preguntarnos entonces es qué aspecto ofrecen las cuestiones de la identidad personal y de su importancia desde este segundo punto de vista, que no sería el parfitiano.

Situados en esta óptica, podremos ver que las razones que tenemos para considerarnos los mismos que nuestros «yoes» futuros no serían en absoluto «metafísicas» (no tendrían que ver con la existencia o no existencia de un *deep further fact*, por ejemplo), sino simplemente prácticas. Es decir, somos personas unificadas en un momento temporal dado porque tenemos que actuar, lo que significa que ocupamos lo que Korsgaard denomina el *deliberative standpoint*: sería *como si* fuésemos algo por encima de nuestros deseos y motivos, el algo que elige a partir de cuál de ellos actuar. Porque la acción racional tiene que ser posible, resulta que somos personas unificadas, del mismo modo que tenemos que reconocernos libres por el mismo motivo; porque tenemos que poder llevar un plan racional de vida resulta que somos personas con una existencia continua a través del tiempo. El de la identidad personal sería como un postulado más de la razón práctica (es la necesidad de actuar la que exige identidad de las personas a través del tiempo, porque esta es condición de posibilidad de aquella, y la acción racional es perfectamente real). El mundo de los experimentos de pensamiento en que nos sumerge Parfit se sitúa del lado reactivo de la tercera antinomia kantiana, es el mundo de las manipulaciones quirúrgicas y de las duplicaciones digitales, el mundo de la tecnología que, como observa Koorsgard, no afecta para nada al lado activo de la tercera antinomia, el de la libertad. Los desarrollos que encontramos en *Razones y personas* permanecerían en todo momento ciegos ante la realidad de la capacidad de actuar, de la *agency*. Porque desde la perspectiva de la capacidad de actuar carece de sentido oponer mi yo presente a los yoes futuros. Se trata de la perspectiva del cuidado de sí, en la que tengo una razón personal para ocuparme de mi «yo»

futuro. De la misma manera, si la «unidad de distribución» tenemos que decir que no es otra que el ser humano, lo es porque el conductor básico de una vida es el ser humano como un todo, y no los segmentos personales en los que nos llevan a pensar muchos experimentos mentales.

Koorsgard concluye su crítica creyendo descubrir, como no podía ser de otra forma, que la radical opción utilitarista de nuestro filósofo es anterior a toda su teoría de la condición de persona, y la condiciona poderosamente, de manera que no resulta en absoluto extraño que su desarrollo termine por confirmarla, como un paradigma que se autoinmuniza. En este extremo coincide esta crítica con la que veremos enseguida de Ricoeur: el Utilitarismo daría el paso decisivo ya en el primer momento, cuando escoge el suceso o acontecimiento como determinación ontológica básica, con lo que iguala la pregunta *¿quién lo hace?* a la cuestión *¿a quién sucede?*, ocultando de este modo desde el mismo punto de partida toda la dimensión práctica en la que, kantianamente entendido, habría que plantear el problema de las personas.

Con lo que se restringe el alcance de las conclusiones parfitianas sobre la identidad personal. El mismo autor se refiere en varias ocasiones al Budismo, conjunto de doctrinas en las que el desenmascaramiento del yo como mera ilusión no se puede separar del decidido abandono del plano de la acción. Sin la acción parece que no tiene mucho sentido la idea misma de yo. Por otra parte, si bien no faltan en la actualidad ensayos de combinar en una ciencia psicológica completa los dos aspectos de la tercera antinomia kantiana, como el de R. Harré con su psicología discursiva, lo indudable es que da la impresión de que una ciencia concebida según los patrones al uso del objetivismo y el naturalismo resulta del todo hostil a la pretensión de dar cabida a la dimensión propiamente activa y espontánea de la realidad humana, que seguiría de este modo *sin tener lugar en el mundo*, como afirmaba el Wittgenstein del *Tractatus*. Con este objetivismo y naturalismo dominantes es bastante solidaria la filosofía de Parfit, caben pocas dudas al respecto, debiendo por nuestra parte buscar su originalidad en el entronque «liberador» de estos aspectos hoy tan presentes en el pensamiento científico y

tecnológico con las posibilidades de un neobudismo que nos descargaría de los agobios de la inmemorial *cura sui*.

También desde los cuarteles hermenéuticos, como es de todos sabido, se han dirigido lúcidos ataques contra lo que serían los sentidos más nucleares de *Razones y personas*. Como ocurre con las grandes polémicas entre tradiciones filosóficas, estaríamos ante las fricciones inevitables de los grandes mundos intelectuales dominantes en la actualidad, fricciones cuya radiografía nos llevaría en este caso al examen de la justificación respectiva de las racionalidades analítica y hermenéutica, algo que desborda con mucho los límites de esta presentación.

Ch. Taylor (1989/1996), por ejemplo, nos remite en su interpretación de la teoría parfitiana de la identidad personal a los avatares de la construcción del individuo moderno, que ha generado una comprensión errónea del yo. De lo que se trataría es de contextualizar el yo de *Razones y personas*, como yo en tanto *objeto* especial (en la medida en que es capaz de autoconciencia) que se ha de conocer, en la tradición empirista de Locke y de Hume. Es decir, un yo *neutro* (que se define fuera de cualquier marco referencial de cuestiones), *puntual* (que se define haciendo abstracción de cualquier inquietud constitutiva) y *desvinculado* (su única propiedad constitutiva es la autoconciencia de un sujeto de control racional). Lo que Taylor le censura a Parfit, concretamente, es su ignorancia del yo como ser que se constituye esencialmente por un cierto modo de preocupación de sí, entendiendo por tal no, desde luego, la preocupación por la cualidad placentera o dolorosa de sus experiencias. Como inserto en la tradición moderna, el filósofo de Oxford permanecería en todo momento al margen de la *comprensión correcta* del yo. Cuidarse de su propia identidad resulta constitutivo para el sujeto humano, lo que entenderíamos en su sentido recto si tenemos en cuenta que la identidad de las personas es una identidad de carácter narrativo.

Así que, si Parfit ha podido insistir tanto en que la identidad no es lo que importa, es porque, como reconoce P. Ricoeur (1990/1996), este filósofo representa el adversario más temible para la tesis, hoy tan corriente, de la identidad narrativa. La clave

de interpretación apuntaría al ocultamiento de la dimensión de la denominada «identidad-*ipse*», que es la que no implicaría nada sobre un supuesto núcleo no cambiante de la personalidad, sino que, al contrario, conllevaría la alteridad. Las conclusiones de Parfit son irresistibles sólo si se admite que la identidad personal quiere decir *mismidad*, identidad-*idem* o simple permanencia en el tiempo. El ocultamiento procede a través de la elección del suceso o acontecimiento, ya lo vimos, como término ontológico de referencia, y la consiguiente eliminación de la que Ricoeur llama «calidad de mío». Semejante elección, puesto que se sitúa en el mismo punto de partida de toda la discusión de la identidad personal a través del tiempo, tiene el efecto inmediato de convertir a la posición reduccionista en la posición por defecto, en la unidad de cuenta, haciendo del No Reduccionismo una tesis que sólo se define por contraste con ella, una tesis por así decir parasitaria. Por definición, el acontecimiento es impersonal, como ese cerebro con cuya consideración exclusiva en los *puzzling cases* se operaría la «neutralización» del cuerpo como *Leib*, lo que los fenomenólogos llaman el cuerpo vivencial o cuerpo propio. También quedaría bien patente el imperialismo del acontecimiento en el vaciado que se hace de todo rasgo de pertenencia propia del horizonte de la memoria personal, en ese concepto chocante pero de importancia crucial para el Reduccionismo que es el concepto de cuasimemoria. El cerebro y la huella mnémica suplantan al cuerpo propio y a la memoria biográfica, con lo que la personalidad quedaría enfocada como verdadera anomalía salvaje en el orden neutro del acontecimiento. Una suplantación de la calidad de mío por la mismidad que se consuma y se consagra en el tipo de Budismo al que acude Parfit para ilustrar lo venerable de sus tesis.

Pero acaba Ricoeur su polémica advirtiendo que la dimensión soslayada sigue ahí siempre, por muy hondo que se la haya enterrado. La *ipseidad* siempre vuelve, como lo reprimido freudiano: Parfit no puede hurtarse «a la tenacidad de los pronombres personales», y, de todos modos, siempre habría que preguntarse quién se pregunta si le importa la identidad personal, como en el caso de Hume nos habíamos preguntado quién busca el yo. Preguntarse por la identidad personal y por su importancia daría testimonio del cuidado de sí que es constitutivo de lo que Ricoeur llama *ipseidad*. Hasta en el desposeimiento de sí característico del neobudismo parfitiano es dable rastrear el señorío de sí de la ipseidad.

El desarrollo del problema de la identidad personal y de sus repercusiones prácticas, tal y como lo encontramos ejemplarmente expuesto en la obra que con este prólogo presentamos en su edición española, no sólo se ha convertido ya en el punto de referencia obligado alrededor del cual se organiza hoy todo el debate metafísico y ético de las personas, sino que nos ha llevado al recíproco cuestionamiento de las tradiciones filosóficas dominantes. Sobre la cuestión de los valores relativos de estos paradigmas de la Filosofía, de su respectiva relevancia para nuestra compleja y problemática vida de hoy, le tocaría decidir a cada uno de nosotros.

Referencias bibliográficas

BLACKBURN, S. (1997), «Has Kant Refuted Parfit?», en Dancy, J. (ed.), 180-202 pp.

DANCY, J. (ed.) (1997), *Reading Parfit*, Oxford: Blackwell.

HARRIS, H. (ed.) (1995), *Identity*, Oxford University Press.

KORSGAARD, Ch. M. (1989), «Personal Identity and the Unity of Agency: A Kantian Response to Parfit», *Philosophy & Public Affairs* (Spring 1989), vol. 18, n.° 2, 101-133 pp.

McDOWELL, J. (1997), «Reductionism and the First Person», en Dancy, J. (ed.), 230-251 pp.

MEYER, L. (2003), «Intergenerational Justice», *The Stanford Encyclopedia of Philosophy (Summer 2003 Edition)*, E. N. Zalta (ed.), URL= http://plato.stanford.edu/archives/sum2003/entries/justice-intergenerational/.

MONTEFIORE, A. (ed.) (1973), *Philosophy and Personal Relations*, London: Routledge and Kegan Paul.

NARVESON, J. (1967), «Utilitarianism and New Generations», *Mind*, 76, 62-72.

PARFIT, D. (1971/1983), «Identidad personal», México: U.N.A.M./ Instituto de Investigaciones Filosóficas (versión castellana de Álvaro Rodríguez Tirado).

— (1971), «On "The Importance of Self-Identity"», *The Journal of Philosophy*, vol. LXVIII, n.º 20, October 21, pp. 667-678.

— (1973), «Later Selves and Moral Principles», en Montefiore, A. (ed.), 137-170 pp.

— (1976/1985), «Lewis, Perry y lo que importa», México: U.N.A.M./Instituto de Investigaciones Filosóficas (versión castellana de Álvaro Rodríguez Tirado).

— (1979/1991), «Prudencia, Moralidad y el Dilema del prisionero», *Excerpta philosophica*, 2, Facultad de Filosofía de la Universidad Complutense (versión castellana de Gilberto Gutiérrez).

— (1986), «Comments», *Ethics*, 96, 832-872 pp.

— (1995a), «Equality or Priority?», Kansas: University of kansas, 1-42 pp.

— (1995b), «The Unimportance of Identity», en Harris, H. (ed.), 13-45 pp.

PERRY, J. (ed.) (1975), *Personal Identity*, Berkeley y Los Angeles: University of California Press.

RICOEUR, P. (1990/1996), *Sí mismo como otro*, Madrid: Siglo XXI (versión castellana de Agustín Neira Calvo).

RORTY, A. O. (ed.) (1976), *The Identities of Persons*, Berkeley: University of California Press.

SCARRE, G. (1996), *Utilitarianism*, London: Routledge.

SHOEMAKER, S. (1984/1997), «Parfit on Identity», en Dancy, J. (ed.), 135-149 pp.

TAYLOR, Ch. (1989/1996), *Fuentes del yo; la construcción de la identidad moderna*, Barcelona: Paidós. (Versión castellana de Ana Lizón.)

TEMKIN, L. S. (1997), «Rethinking the Good, Moral Ideals and the Nature of Practical Reasoning», en Dancy, J. (ed.), 290-346 pp.

[MARIANO RODRÍGUEZ]

RAZONES Y PERSONAS

A mis padres
Drs. Jessie y Norman Parfit.
Y a mis hermanas
Theodora y Joanna.

«Por fin el horizonte se nos muestra libre otra vez, aunque desde luego no esté claro; por fin nuestros barcos pueden aventurarse a salir otra vez, aventurarse a afrontar cualquier peligro. Todo el atrevimiento del amante del conocimiento se permite otra vez. El mar, *nuestro* mar, se extiende abierto otra vez. Quizás nunca haya habido un "mar abierto" como éste.»

Nietzsche*

* *La gaya ciencia,* V, 343. (N. del t.)

AGRADECIMIENTOS

Hace dieciséis años viajaba yo a Madrid con Gareth Evans. Por entonces tenía la esperanza de llegar a ser filósofo, de manera que cuando conducíamos por Francia le expuse mis ideas. Sus críticas me pusieron al borde de la desesperación. Pero antes de llegar a España me di cuenta de que era casi tan crítico con sus propias ideas. Como tantos otros, debo mucho a la intensidad de su amor a la verdad y a su extraordinaria vitalidad. Dejo constancia de esta deuda antes que de ninguna otra porque murió con 34 años.

Debo mucho a mis primeros maestros: Sir Peter Strawson, Sir Alfred Ayer, David Pears y Richard Hare. Desde entonces he aprendido de muchos. Discutiendo, sobre todo de Thomas Nagel, Ronald Dworkin, Tim Scalon, Amartya Sen, Jonathan Glover, James Griffin, Ann Davis, Jefferson McMahan y Donald Regan. He aprendido mucho más de la lectura de los escritos de éstos y de otros muchos. Parte de mis deudas las reconozco en las notas finales de este libro. Pero estoy seguro de que, a causa de mi débil memoria y por no haber tomado las notas apropiadas, este libro presenta como ideas mías muchas tesis o argumentos que en realidad debo atribuir a alguna fuente exterior a mí. Estas fuentes olvidadas, si llegasen a leer este libro, se sentirían dolidas con toda razón. Aunque debe-

rían mencionarse en las notas finales, espero que la mayoría se mencionen al menos en la Bibliografía.

Algunas personas me ayudaron a escribir este libro. Antes de que falleciera hace dos años, John Mackie escribió unos comentarios extremadamente útiles sobre mi trabajo anterior. En los últimos meses he recibido muchos comentarios sobre un borrador de este libro —tantos que no he tenido tiempo de hacer todas las revisiones necesarias—. Esta es una lista dispuesta al azar de los que me han ayudado de esta manera: Jonathan Glover, Sir Peter Strawson. John McDowell, Susan Hurley, Paul Seabright, John Vickers, Hywel Lewis, Judith Thomson, Samuel Scheffler, Martin Hollis, Thomas Nagel, Robert Nozick, Richard Lindley, Gilbert Harman, Christopher Peacocke, Peter Railton, Anette Baier, Kurt Baier, Richard Swinburne, Michael Tooley, Mark Sainsbury, Wayne Sumner, Jim Stone, Dalie Jamieson, Eric Rakowski, James Griffin, Gregory Kavka, Thomas Hurka, Geoffrey Madell, Ralph Walker, Bradford Hooker, Douglas Maclean, Graeme Forbes, Bimal Matilal, Nicholas Dent, Robert Goodin, Andrew Brennan, John Kenyon, James Fishkin, Robert Elliott, Arnold Levison, Simon Blackburn, Ronald Dworkin, Amartya Sen, Peter Unger, Peter Singer, Jennifer Whiting, Michael Smith, David Lyons, Milton Wachsberg, William Ewald, Galen Strawson, Gordon Cornwall, Richard Sikora, Partha Dasgupta, Dr. Jessie Parfit y Dr. Charles Whitty.

He aprendido algo de cada uno de los nombrados, y de algunos aprendí mucho. De unas cuantas personas aprendí tanto que quiero agradecérselo especialmente. Jonathan Bennett me envió comentarios muy provechosos sobre la mitad de mi borrador. Bernard Williams me mandó unos comentarios extremadamente provechosos sobre un borrador de la Tercera Parte. Otras seis personas me remitieron comentarios muy útiles acerca de borradores de todo el libro. Cuatro de ellas fueron John Leslie, Micharl Woodford, Larry Temkin y Donald Regan. Y todavía aprendí más de las otras dos. John Broome fue *profesor invitado* en mi *College* durante el año académico a cuya finalización escribo estas palabras. Tanto en comentarios escritos como en muchas discusiones solucionó un buen número de mis problemas y propició muchas grandes mejoras. Si mencionase en las notas finales cada pasaje que debo a John Broome, habría por lo menos treinta de tales notas. Puesto que las diferentes disciplinas académicas se van separando cada vez más de sus vecinas, resulta reconfortante encontrarse con que un economista puede ser tan buen filósofo en sus ratos libres. Pero la persona de la que he aprendido más es Shelly Kagan. Sus comentarios, extraordinariamente agudos y penetrantes, llegaron a ocupar tanto como la mitad de mi borrador, y muchas de sus sugerencias se han impreso, con cambios mínimos, en este libro. Si su coautoría fuese mencionada en las notas finales, habría al menos unas sesenta de esas notas.

Escribo estas palabras el día antes de que este texto vaya a la imprenta. Como recibí tantas buenas objeciones y comentarios, no lo hubiera podido revisar y confeccionar a tiempo sin la ayuda de otras personas. Me ayudó Patricia Morison, me ayudaron mucho Susan Hurley y William Ewald. Jefferson y Sally McMahan me ahorraron muchos días de trabajo clasificando notas, comprobando referencias y compilando la Bibliografía. Este libro se ha impreso por el procedimiento de la fotocomposición. Dada mi lentitud a la hora de hacer las revisiones necesarias, las cuatro personas que realizaron la copia han tenido que trabajar en horas extraordinarias, hasta bien entrada la noche, y todo ello sin quejarse. Estas personas generosas son Angela Blackburn, Jane Nunns, Paul Salotti, y, la más generosa de todas, Catherine Griffin.

Estoy agradecido por la ayuda de todos los que he citado. A los que he mencionado en los dos últimos párrafos quiero dejarles aquí constancia de la mayor de mis gratitudes. Este libro tiene un autor, pero es realmente el producto colectivo de todas esas personas.

Finalmente, quiero expresar mi gratitud a una entidad que no es una persona: el *All Souls College*. Si yo no hubiera tenido el extraordinario privilegio de ser *profesor galardonado* y después *miembro del equipo investigador* de este *College* durante los últimos dieciséis años, este libro, con toda certeza, no existiría.

All Souls College, Oxford D.A.P.
12 de septiembre de 1983

INTRODUCCIÓN

Como mi gato, a menudo simplemente hago lo que quiero hacer. Entonces no uso una habilidad que sólo tienen las personas. Sabemos que hay razones para actuar, y que algunas razones son mejores o más fuertes que otras. Uno de los temas principales de este libro es un conjunto de preguntas sobre aquello que tenemos razones para hacer. Discutiré varias teorías. Algunas de ellas son morales, otras son teorías acerca de la racionalidad.

Somos personas individuales. Yo tengo que vivir mi vida y tú la tuya. ¿Qué implican estos hechos? ¿Qué es lo que hace de mí la misma persona durante mi vida, y una persona diferente de ti? ¿Y qué importancia tienen estos hechos? ¿Qué importancia tiene la unidad de cada vida, y la distinción entre diferentes vidas, y entre diferentes personas? Estas preguntas son el otro tema principal de este libro.

Mis dos temas, razones y personas, tienen conexiones estrechas. Creo que la mayoría de nosotros tiene falsas creencias sobre nuestra propia naturaleza y sobre nuestra identidad a través del tiempo, y que, cuando vemos la verdad, debemos cambiar algunas de nuestras creencias acerca de lo que tenemos razones para hacer. Debemos revisar nuestras teorías morales y nuestras creencias acer-

ca de la racionalidad. En las dos primeras partes del libro doy otros argumentos en favor de conclusiones similares.

No describiré, para no adelantarme, estos argumentos y conclusiones. El Índice aporta un sumario. El libro es largo, y a veces complicado. Por eso he separado mis argumentos en 154 partes, dándole a cada una de ellas un título descriptivo. Espero que esto haga los argumentos más fáciles de seguir, y muestre el contenido del libro más claramente de lo que lo haría un Índice de Temas. Si no hubiera vuelto a poner en orden los argumentos en estas partes separadas, tal Índice habría estado demasiado lleno de referencias como para ser de alguna utilidad.

Muchas introducciones a libros de este tipo intentan explicar los conceptos centrales que se usan. Como hubiera llevado al menos un libro el dar una explicación útil, no perderé el tiempo haciendo menos que eso. Mis conceptos centrales son pocos. Tenemos *razones para actuar*. *Debemos* actuar de determinadas maneras, y algunas maneras de actuar son *moralmente incorrectas*. Ciertos resultados son *buenos* o *malos*, en un sentido que tiene relevancia moral: es malo, por ejemplo, que una persona se quede paralizada, y si podemos debemos evitarlo. La mayoría de nosotros comprende mis tres últimas frases lo bastante bien como para comprender mis argumentos. También utilizaré el concepto de lo que se incluye en el *propio interés* de alguien, o lo que sería *mejor para esta persona*. Discuto esto brevemente en el Apéndice I. Mi último concepto central es el de *persona*. La mayoría de nosotros piensa que sabemos lo que son las personas. La Tercera Parte sostiene que no lo sabemos.

Muchas introducciones también intentan explicar cómo podemos esperar hacer progresos al discutir la moralidad. Ya que la mejor explicación nos la daría el *progresar*, esa es la única explicación que trataré de dar.

Strawson describe dos clases de filosofía, la descriptiva y la revisionista. La filosofía descriptiva proporciona razones para lo que nosotros asumimos instintivamente, y explica y justifica el núcleo central inamovible de nuestros conceptos acerca de nosotros mismos y del mundo en que vivimos. Tengo un gran respeto por la filosofía descriptiva. Pero, temperamentalmente, soy un revi-

sionista. Excepto en mi árido Capítulo 1, donde no puedo evitar repetir lo que se ha demostrado que es verdadero, intento desafiar lo que damos por establecido. Los filósofos no sólo deberían interpretar nuestras creencias. Cuando sean falsas, deberían *cambiarlas*.

Nota añadida en 1985: En esta reimpresión he hecho varias correcciones. Así, le he retirado mi apoyo al Criterio Psicológico Amplio de identidad personal (pp. 382 y 383 entre otras), puesto que entra en conflicto con mi idea de que no deberíamos intentar decidir entre los diferentes criterios. He sustituido "malo" por "peor" en la Tesis C (p. 621) y en varias frases posteriores. Otras correcciones sustantivas mínimas incluyen las que se realizaron en las Notas 15, a la Parte I, y 83, a la Parte III, y en las páginas 158/líneas 7-9, 160/6-9, 387/30-32, 544/23-24, y 642/35-37 y 643/12-14. Por algunas de estas correcciones estoy agradecido a los colaboradores de *Ética*, julio 1986.

Nota añadida en 1987: En esta reimpresión he ampliado mi definición de Reduccionismo (p. 385-386), y he eliminado una circularidad aparente de los Criterios Físico y Psicológico (pp. 377 y 381). Otras correcciones sustantivas mínimas se realizaron en las páginas 167/líneas 2-4, 168/13-27, 181/28-30 y 182/1-8, 183/1-8, 543/15-16 y la nota 59 a la Parte I. (También hice varias correcciones o revisiones estilísticas.)

Primera parte

TEORÍAS CONTRAPRODUCENTES

1

TEORÍAS QUE SON INDIRECTAMENTE CONTRAPRODUCENTES

¿Para hacer qué cosas tenemos más razones? Varias teorías contestan esta a pregunta. Algunas de ellas son teorías morales; otras, teorías de la racionalidad. Cuando las aplicamos a algunas de nuestras decisiones, diferentes teorías nos dan diferentes respuestas. Tenemos entonces que tratar de decidir cuál es la mejor teoría.

Los argumentos sobre estas teorías son de muchas clases. Un argumento consiste en que una teoría *es contraproducente*. Este argumento, sólo él, no necesita suposiciones. Establece que una teoría falla incluso en sus propios términos, y así se condena a sí misma.

La primera parte de este libro discute lo que este argumento consigue. Como explicaré más adelante, todas las teorías mejor conocidas son en cierto sentido contraproducentes. ¿Qué demuestra esto? En algunos casos, nada. En otros, lo que se demuestra es que una teoría tiene que ser desarrollada o ampliada en mayor medida. Y en otros casos lo que se demuestra es que una teoría tiene que ser o rechazada o revisada. Esto es lo que se demuestra en relación con las teorías morales que la mayoría de nosotros acepta.

Comenzaré con el caso mejor conocido.

1. La Teoría del Propio Interés

Podemos describir todas las teorías diciendo lo que nos indican que tratemos de lograr. Según todas las teorías morales, debemos tratar de actuar moralmente. Según todas las teorías de la racionalidad, debemos tratar de actuar racionalmente. Llamemos a estos nuestros fines *formales*. Diferentes teorías morales, y diferentes teorías de la racionalidad, nos dan diferentes fines *sustantivos*.

Por «fin» querré significar «fin sustantivo». Este uso de fin es amplio. Puede describir teorías morales que se ocupan no de metas morales, sino de derechos o deberes. Supongamos que, según cierta teoría, cinco clases de actos están totalmente prohibidos. Esta teoría nos da a cada uno de nosotros el fin de no actuar nunca de ninguna de esas cinco maneras.

Discutiré en primer lugar la *Teoría del Propio Interés*, o PI. Esta es una teoría de la racionalidad. PI da a cada persona este fin: los resultados que serían los mejores para ella misma, y que hagan que su vida marche, para ella misma, lo mejor posible.

Para aplicar PI, tenemos que preguntar qué es lo que conseguiría este fin del mejor modo posible. Las respuestas a esta pregunta las denomino *teorías del propio interés*. Como explica el Apéndice I, hay tres teorías plausibles.

Según la *Teoría Hedonista*, lo que sería lo mejor para alguien es lo que le diera mayor felicidad. Las diferentes versiones de esta teoría hacen diferentes afirmaciones sobre lo que la felicidad lleva consigo, y cómo se la debería medir.

Según la *Teoría de la Realización de Deseos*, lo que sería lo mejor para alguien es lo que realizara mejor sus deseos a lo largo de su vida. Aquí de nuevo hay versiones diferentes de esta teoría. Así, la *Teoría del Éxito* apela sólo a los deseos de una persona sobre su propia vida.

Según la *Teoría de la Lista Objetiva*, ciertas cosas son buenas o malas para nosotros, aun cuando no quisiéramos tener las buenas ni evitar las malas. Aquí de nuevo hay diferentes versiones. Las cosas buenas podrían incluir el desarrollo de las propias habilidades, del conocimiento y de la conciencia de la verdadera belleza. Las cosas malas podrían incluir el placer sádico, el ser engañado y la pérdida de la libertad o de la dignidad.

Estas tres teorías se superponen parcialmente. Según todas ellas, la felicidad y el placer son, como mínimo, parte de lo que hace que nuestras vidas marchen mejor para nosotros, y el sufrimiento y el dolor son, como mínimo, parte de lo que hace que nuestras vidas vayan peor. Estas afirmaciones las haría cualquier Teoría de la Lista Objetiva que fuese plausible. Y las tres están implicadas por todas las versiones de la Teoría de la Realización de Deseos. Al parecer de todas las teorías, pues, la Teoría Hedonista es, como mínimo, parte de la verdad. Para ahorrar palabras, esta será en ocasiones la única parte que discutiré.

Todas estas teorías afirman asimismo que, al decidir lo que sería lo mejor para alguien, deberíamos dar igual importancia a todas las partes del futuro de esta persona. Los sucesos posteriores pueden ser menos predecibles, y un suceso predecible debería contar menos si es menos probable que ocurra. Pero no debería contar menos simplemente porque, si ocurre, vaya a ocurrir más tarde.

Haría falta como mínimo un libro para decidir entre las diferentes teorías del propio interés. Este libro discute algunas de las diferencias entre estas teorías, pero no intenta decidir entre ellas. Gran parte de este libro discute la teoría del Propio Interés. Como he dicho, esta no es una de las teorías acerca del propio interés. Es una teoría de la racionalidad. Podemos discutir PI sin decidir entre las diferentes teorías acerca del propio interés. Podemos formular tesis que serían verdaderas según todas estas teorías.

Será útil denominar a algunos fines *últimos*. Otros son *instrumentales*, simples medios para la consecución de algún fin último. Así, para todos a excepción de los avaros, ser rico no es un fin último. Ahora puedo reformular la *tesis central* de PI. Es esta

> (PI1) Hay para cada persona un fin último supremamente racional: que su vida marche, para ella, de la mejor manera posible.

Como después veremos, PI hace otras varias afirmaciones.

Hay varias objeciones a PI. Algunas de ellas las trato en las Partes Segunda y Tercera. En lo que sigue voy a discutir la objeción de que, al igual que otras teorías, PI es contraproducente.

2. Cómo puede ser PI indirectamente contraproducente

Si llamamos T a cierta teoría, llamaremos a los fines que ella nos asigna *nuestros fines T-dados*. Y diremos que T es

> *individualmente contraproducente de manera indirecta* cuando ocurre que, si alguien intenta lograr sus fines T-dados, estos fines serán peor logrados, contempladas las cosas en su conjunto.

Según esta definición, no nos limitamos a preguntar si una teoría es contraproducente. Lo que preguntamos es si es contraproducente cuando se aplica a determinadas personas durante ciertos periodos.

Mi fin PI-dado es que mi vida vaya, para mí, lo mejor posible. Puede ser verdad que, si yo trato de hacer aquello que será lo mejor para mí, esto será peor para mí. Hay dos tipos de casos:

(a) Mis intentos pueden a menudo fallar. Con frecuencia puedo hacer lo que para mí será peor que alguna otra cosa que yo podría haber hecho.
(b) Aunque yo nunca haga lo que será peor para mí de los actos que me son posibles, puede ser peor para mí si me guío puramente por mi propio interés. Podría favorecerme el tener una disposición distinta.

En casos de tipo (a), los efectos negativos proceden de lo que hago. Supongamos que robara siempre que creyera que no me van a coger. En muchas ocasiones puede ocurrir que me cojan y me castiguen. Aún en los términos del propio interés, la honestidad puede por tanto ser la mejor política para mí. No vale la pena discutir estos casos. Si así es como PI es contraproducente, esto no es objeción para PI. Porque PI es aquí contraproducente sólo a causa de mi

incompetencia al intentar seguir PI. Esto es un defecto, pero no de PI, sino mío. Podríamos objetar a cierta teoría que es demasiado difícil de seguir. Pero esto no es cierto de PI.

Los casos que vale la pena discutir son los del tipo (b). En estos casos sucede que será peor para mí si me conduzco puramente en los términos del propio interés, aunque tuviera éxito en no hacer nunca lo que será peor para mí. Los efectos negativos no proceden de lo que hago sino de mi disposición, o sea, del hecho de que me guío puramente por el propio interés.

¿Qué conlleva este hecho? Yo podría seguir puramente el propio interés sin ser puramente egoísta. Supongamos que quiero a mi familia y a mis amigos. Según todas las teorías del propio interés, mi amor hacia esas personas afecta a lo que se incluye en mis intereses. Gran parte de mi felicidad viene de saber que los que quiero son felices, y de ayudar a promover esta su felicidad. Según la Teoría de la Realización de Deseos, será mejor para mí si, como yo deseo, las cosas les van bien a los que quiero. Lo que será lo mejor para mí puede, de estas y de otras maneras, coincidir en gran medida con lo que será lo mejor para aquellos a quienes quiero. Pero en ciertos casos, lo que será mejor para mí será peor para los que quiero. Y yo me guiaré por mi propio interés si, en todos estos casos, hago lo que será mejor para mí.

Puede pensarse que, si me guío por mi propio interés, *siempre* intentaré hacer todo lo que será mejor para mí. Pero con frecuencia actúo de una de dos maneras, creyendo que ninguna sería la mejor para mí. En estos casos no estoy intentando hacer lo que será lo mejor para mí: estoy actuando a partir de un deseo más particular. Y esto puede ser verdad aunque esté haciendo lo que sé que será lo mejor para mí. Supongamos que sé que, si te ayudo, esto será lo mejor para mí. Puedo ayudarte porque te quiero, no porque quiera hacer lo que será lo mejor para mí. Al describir lo que sería para mí guiarme por el propio interés, basta con afirmar que, mientras que a menudo yo actúo sobre la base de otros deseos, yo *nunca hago lo que creo que será peor para mí*. Si esto es verdad, será más claro decir, no que *me guío por el propio interés*, sino que *nunca soy abnegado*.

Redescribiré a continuación la forma tan interesante en que, para un individuo dado, PI puede ser indirectamente contraprodu-

cente. Esto sucede en el momento en que, si alguien no es nunca abnegado, esto es peor para él que si tuviera una disposición diferente. Aunque alguien tenga éxito en no hacer nunca lo que sería peor para él, puede ser peor para él que no sea nunca abnegado. Podría ser mejor para él tener alguna otra disposición. Así que en ocasiones podría hacer lo que fuese peor para él. Pero los costes que le supondrían a él actuar de este modo podrían ser menores que los beneficios de tener esta otra disposición.

Estas afirmaciones pueden ser verdaderas según todas las teorías del propio interés. Hace mucho tiempo que los hedonistas saben que cuando perseguimos la felicidad es más difícil de conseguir. Si mi deseo más fuerte es ser feliz, puedo ser menos feliz de lo que sería si tuviese otros deseos que fueran más fuertes. Así que yo podría ser más feliz si mi deseo más fuerte fuese que alguien diferente de mí fuese feliz.

Aquí tenemos otro ejemplo. *Kate* es escritora. Su más ferviente deseo es que sus libros sean los mejores posibles. Como se preocupa tanto de la calidad de los libros que escribe, encuentra su trabajo muy gratificante. Si el deseo que tiene de escribir buenos libros fuese mucho más débil, encontraría su trabajo aburrido. Ella lo sabe, y acepta la Teoría Hedonista del propio interés. Cree por consiguiente que es mejor para ella que su más ferviente deseo sea que sus libros sean tan buenos como sea posible. Sin embargo, a causa de la fuerza de este deseo, a menudo trabaja demasiado duro, y por periodos de tiempo excesivos, hasta el punto de que acaba completamente agotada y se queda muy deprimida durante largas temporadas.

Supongamos que Kate creyera, de forma correcta, que si trabajara con menos intensidad sus libros serían algo peores, pero ella sería más feliz. Entonces encontraría su trabajo igual de gratificante, pero evitaría esas severas depresiones. Kate, por tanto, cree que, cuando trabaja con tanta dureza, está haciendo lo que es peor para ella. Pero ¿cómo podría llegar a suceder que ella nunca actuara de ese modo? Ella nunca trabajaría de ese modo sólo con la condición de que tuviera un deseo mucho más débil de que sus libros fueran los mejores posibles. Y esto sería aún peor para ella porque entonces encontraría su trabajo aburrido. Según la Teoría Hedonista, sería peor para Kate que nunca fuese abnegada [1].

Supongamos que no aceptamos la Teoría Hedonista del propio interés, sino la Teoría de la Realización de Deseos. Podemos entonces negar que, en este ejemplo, Kate esté haciendo lo que es peor para ella. Su deseo más fuerte es que sus libros sean tan buenos como sea posible. Trabajando con tanta dureza, aunque acabe exhausta y deprimida, consigue que sus libros sean algo mejores. Con ello está haciendo que su deseo más potente se realice mejor. Así que, según nuestra teoría del propio interés, esto puede ser mejor para ella.

Si no somos hedonistas, necesitamos un ejemplo diferente. Supongamos que conduzco mi coche a medianoche por un paraje desértico. El coche se estropea. Tú eres un desconocido, y el único conductor que pasa por allí. Me las arreglo para que te detengas y te ofrezco una gran recompensa si me sacas del apuro y me rescatas. No puedo recompensarte ahora, pero prometo hacerlo cuando lleguemos a mi casa. Supongamos también que yo soy *transparente*, incapaz de engañar a nadie. Soy incapaz de mentir de manera convincente. Ya sea que me sonrojo, ya sea por causa del tono de mi voz, siempre acabo delatándome. Y supongamos, por último, que yo sé que nunca soy abnegado. Si me llevas en tu coche hasta mi casa, sería peor para mí si te diera la recompensa prometida. Puesto que yo sé que nunca voy a hacer lo que será peor para mí, sé que romperé mi promesa. Dada mi incapacidad de mentir convincentemente, tú también lo sabes. No crees en mi promesa, y por eso me dejas tirado en el desierto. Lo cual me sucede porque yo nunca soy abnegado. Hubiera sido mejor para mí haber sido *digno de confianza*, haber estado dispuesto a cumplir mis promesas aun cuando hacerlo así hubiese sido peor para mí. Porque entonces me habrías rescatado.

(Se puede objetar a esto que, aunque nunca vaya a ser abnegado, yo podría decidir cumplir mi promesa, porque tomar esta decisión sería mejor para mí. Si yo decidiera cumplir mi promesa, tú confiarías en mí y me rescatarías. Pero se puede responder a esta

[1] Para otro ejemplo, véase el comienzo de Adams (2)*.
* Ver en bibliografía, pp. 885-901.

objeción. Yo sé que, después de haberme llevado a casa en tu coche, sería peor para mí darte la recompensa prometida. Si yo sé que nunca actúo abnegadamente, sé que no cumpliré mi promesa. Y, si yo sé esto, no puedo decidir cumplir mi promesa. Porque no puedo decidir hacer lo que sé que no voy a hacer. Si yo *puedo* decidir cumplir mi promesa, esto tiene que ser porque yo no creo que nunca vaya a actuar abnegadamente. Podemos añadir el supuesto de que yo no creería esto a no ser que fuese verdadero. Sería entonces verdadero que es para mi peor que yo nunca fuese y nunca siguiese siendo abnegado. Sería para mí mejor ser digno de confianza.)

He descrito dos modos en que sería peor para una persona no ser abnegada. Hay otras muchas maneras en que esto puede verificarse. Probablemente ocurra así con la mayoría de la gente, durante la mayor parte de sus vidas. Cuando aplicamos la teoría del Propio Interés a estas personas, ésta se nos revela como lo que yo llamo indirecta e individualmente contraproducente. ¿Esto hace que PI falle en sus propios términos? ¿Se condena PI a sí misma? Esto depende de si PI dice a estas personas que nunca se comporten abnegadamente.

3. ¿Nos dice PI que nunca seamos abnegados?

Puede parecer obvio que PI dice a todos que nunca sean abnegados. Pero, tal y como la he descrito hasta este momento, PI afirma sólo que, para cada persona, hay sólo un fin último supremamente racional: que su vida marche, para ella, lo mejor posible.

Cuando se aplica a actos, PI establece tanto que

> (PI2) Lo que cada uno de nosotros tiene [2] más razón para hacer es todo aquello que sería mejor para sí mismo, como que
> (PI3) Es irracional para cualquiera hacer lo que cree que sería peor para sí mismo.

[2] Si *hay* una razón para que alguien haga X, diré que esta persona *tiene* una razón para hacer X. Según este uso de las palabras, podemos tener razones para actuar de las cuales no somos conscientes.

PI también tiene que hacer afirmaciones acerca de lo que debiéramos hacer cuando no podemos predecir los efectos de nuestros actos. Podemos ignorar los casos de *incertidumbre*, en los que no disponemos de creencias sobre las probabilidades de diferentes efectos. En los casos de *riesgo*, donde sí que tenemos tales creencias, PI afirma que

> (PI4) Lo que sería racional hacer para cualquiera es aquello que le fuera a reportar el mayor beneficio *esperado*.

Para calcular el beneficio esperado de cierto acto, sumamos los beneficios posibles y restamos los costes posibles, con cada beneficio o coste multiplicado por la posibilidad de que el acto lo produzca. Así, si determinado acto tiene una posibilidad de nueve entre diez de reportarme determinado beneficio B, y una posibilidad de uno entre diez de hacerme perder cierto beneficio que sería dos veces B, el beneficio esperado es B x 9/10 - 2B x 1/10, o siete décimos de B.

¿Qué debe establecer PI sobre la racionalidad de los deseos y las disposiciones? Puesto que PI afirma que, para cada persona, hay un fin último supremamente racional, PI debe claramente establecer que el deseo supremamente racional es el deseo de que este fin sea logrado. PI debe establecer que

> (PI5) El deseo supremamente racional consiste en que la vida le vaya a cada cual lo mejor posible.

De forma similar, PI debe establecer que

> (PI6) La disposición supremamente racional es la de aquel que nunca es abnegado.

Si alguien no es nunca abnegado, aunque a veces actúe sobre la base de otros deseos nunca actuará contra el deseo supremamente racional. Nunca hará lo que crea que es peor para él.

Para ahorrarnos palabras llamaremos a deseos y disposiciones *motivos*. Hay modos en que, a lo largo del tiempo, podemos hacer

que nuestros motivos cambien. Podemos desarrollar hábitos. Si actuamos de modos con los que ahora no disfrutamos, puede llegar el momento en que disfrutemos de ellos. Si cambiamos de trabajo, o de residencia, o leemos ciertos libros, o tenemos hijos, esto puede causar cambios predecibles en nuestros motivos. Y hay muchas otras maneras en que podemos causar tales cambios.

Según (PI2), lo que toda persona tiene más razón de hacer es determinarse a tener, o permitirse mantener, cualquiera de los *mejores conjuntos posibles de motivos, en términos del propio interés.* Estos son aquellos conjuntos de motivos de los cuales es verdadero lo que sigue. No hay otro posible conjunto de motivos del cual sea verdadero que, si esta persona tuviese estos motivos, esto sería mejor para ella. Por «posible» quiero decir «causalmente posible, dados los hechos generales acerca de la naturaleza humana y los hechos particulares acerca de la naturaleza de esta persona». Muchas veces sería difícil saber si algún conjunto de motivos sería causalmente posible para alguien, o si sería uno de los mejores conjuntos para esta persona en términos del propio interés. Pero podemos ignorar estas dificultades. Hay muchos casos en que alguien sabe que sería mejor para él que hubiera algún cambio en sus motivos. Y en muchos de estos casos tal persona sabe que, de una de las maneras descritas arriba, ella podría causar este cambio. (PI3) implica que sería irracional para esta persona no causar este cambio.

Afirmaciones similares se aplicarían a nuestras emociones, creencias, habilidades, el color de nuestro pelo, el lugar donde vivimos y todo lo demás que podríamos cambiar. Lo que cada uno de nosotros tiene más razón para hacer es producir cualquier cambio que sea mejor para sí mismo. Si alguien cree que podría producir tal cambio, sería irracional para él no hacerlo así.

Ahora podemos volver a mi pregunta anterior. Estamos discutiendo a las personas de las que es verdadero que, si no son nunca abnegadas, esto sería para ellas peor que si tuvieran alguna otra disposición. Supongamos que estas personas saben que esto es verdadero. ¿Les dice PI que no sean nunca abnegadas?

PI afirma lo siguiente. Si una tal persona nunca fuese abnegada, tendría la disposición que es máximamente racional. Pero sería irracional para esta persona determinarse a tener, o a mantener, esta disposición. Sería para ella racional determinarse a tener, o a mantener, la otra disposición, desde el momento en que esto sería mejor para ella.

Estas afirmaciones puede que parezcan dar respuestas contradictorias a mi pregunta. Pueden parecer decirle a esta persona las dos cosas, que sea y que no sea nunca abnegada.

Esto malinterpreta PI. Cuando PI afirma que una disposición es máximamente racional, no nos dice que *tengamos* esta disposición. Recordemos la distinción entre fines formales y fines sustantivos. Como todas las teorías sobre la racionalidad, PI le da a todo el mundo este fin formal: sé racional y actúa racionalmente. Lo que distingue a las diferentes teorías es que nos dan diferentes fines sustantivos. En su afirmación central, (PI1), PI le da a toda persona un fin sustantivo: que su vida vaya, para ella, lo mejor posible. ¿Acaso PI le da a toda persona *otro* fin sustantivo: sé racional y actúa racionalmente? No. Según PI, nuestro fin formal no es un fin sustantivo.

Se podría pensar que, al hacer estas afirmaciones, no he descrito la mejor versión de la teoría del Propio Interés. Pero esta es la versión que sería aceptada por la mayoría de los que creen en esta teoría. La mayoría de estas personas acepta (PI2) y (PI3). Supongamos que yo sé que lo mejor para mí es hacerme irracional. Pronto describiré un caso en el que esto casi podría ser verdadero con toda certeza. Si esto es verdadero, (PI2) implica que sería irracional para mí *no* hacerlo así. Estas afirmaciones no me fijan como fin sustantivo el ser racional.

¿Implica esto que, para PI, ser racional es un simple medio? Esto depende de cuál sea la mejor teoría sobre el propio interés. Para la Teoría Hedonista, PI le da a toda persona este fin sustantivo: la mayor felicidad posible para sí misma. Ser racional no es una parte esencial de *este* fin. Es un mero medio. También lo es actuar racionalmente, y tener deseos o disposiciones racionales. Consideremos a continuación la Teoría de la Lista Objetiva. Según algunas versiones de esta teoría, ser racional es una de las cosas que es buena para toda persona, cualquiera que sean los efectos que ello pueda

tener. Si esto es así, ser racional no es un mero medio, sino parte del fin sustantivo que PI da a toda persona. Lo mismo sería verdadero, para la Teoría de la Realización de Deseos, en el caso de esas personas que quieren ser racionales, sean los que sean los efectos que ello pueda tener.

Se podría objetar: «Supongamos que aceptamos la Teoría Hedonista. PI entonces nos dice que ser racional es un mero medio. Si esto es así, ¿por qué deberíamos intentar ser racionales? ¿Por qué deberíamos querer saber lo que tenemos más razones para hacer? Si aceptamos lo que PI establece, y creemos que ser racional es un mero medio, dejaremos de estar interesados en las preguntas que PI afirma responder. Esto tiene que ser una objeción contra PI. Una teoría aceptable de la racionalidad no puede afirmar que ser racional es un mero medio».

Podríamos responder: «Una teoría sería inaceptable si afirmara que ser racional no importa. Pero no es esto lo que PI afirma. Supongamos que me agarro a una roca como un mero medio de escapar a la muerte. Aunque mi acto es un mero medio, importa mucho. Lo mismo puede ser verdadero acerca de ser racional». Puede que esto no conteste completamente a esta objeción. Como veremos, hay una objeción similar a ciertas teorías morales. Para ahorrar palabras, discutiré estas objeciones al mismo tiempo. Esta discusión tiene lugar en la Sección 19.

Ahora puedo explicar una observación que hice arriba. Según PI, la disposición que es máximamente racional es la de alguien que nunca es abnegado. Yo escribí que, al hacer esta afirmación, PI no nos dice que tengamos esta disposición. PI le da a toda persona un fin sustantivo: que su vida marche, para ella, lo mejor posible. Para algunas teorías del propio interés, ser racional sería, para ciertas personas, parte de este fin. Pero esto sólo sería porque, al igual que ser feliz, ser racional es parte de lo que hace que nuestras vidas marchen mejor. Ser racional no es, *como tal*, un fin sustantivo. Ni tampoco lo es tener la disposición máximamente racional.

En el caso de ciertas personas, según PI, ser racional *no* sería parte de lo que hace que sus vidas marchen mejor. Estas son las personas que estoy discutiendo. Es verdadero de estas personas que, si

nunca fuesen abnegadas, esto sería peor para ellas que tener alguna otra disposición. Puesto que esto es verdadero, no ser nunca abnegado *no* sería parte del fin que PI les da a estas personas. PI no les dice a estas personas que tengan lo que PI afirma ser la disposición máximamente racional: la que es propia del que nunca es abnegado. Y, si ellas pueden cambiar su disposición, PI les dice a estas personas, si ellas pueden, que *no* se propongan no ser nunca abnegadas. Porque sería mejor para estas personas que tuviesen alguna otra disposición, PI les dice que se determinen a sí mismas a tener, o a mantener, esta otra disposición. Si ellas saben que podrían actuar de una de estas maneras, PI afirma que sería irracional para ellas no hacerlo así. Sería irracional para ellas determinarse a sí mismas a no ser, o permitirse a sí mismas no seguir siendo, nunca abnegadas.

4. Por qué PI no falla en sus propios términos

Estas afirmaciones contestan la otra pregunta que hice. Cuando PI se aplica a estas personas es lo que llamo indirectamente contraproducente. ¿Esto hace que PI falle en sus propios términos? ¿Se condena PI a sí misma?

La respuesta es No. PI es indirectamente contraproducente porque sería peor para estas personas no ser nunca abnegadas. Pero PI *no* les dice a estas personas que nunca sean abnegadas. Les dice que lo sean, si es que pueden. Si estas personas nunca son abnegadas, esto es peor para ellas. Esto es un mal efecto, en términos de PI. Pero este mal efecto no es el resultado ni de hacer lo que PI les dice que hagan, ni de tener una disposición que PI les dice que tengan. Por consiguiente, PI no falla en sus propios términos.

Puede objetarse: «Este mal efecto puede ser el resultado de que estas personas *crean* en PI. Porque si creen en PI, creen que sería irracional para ellas hacer lo que creen que será peor para ellas. Puede ser verdadero que, si ellas creen que es irracional actuar de este modo, nunca lo hagan. Si nunca actúan de este modo, nunca son abnegadas. Supongamos que, en uno de los modos que tú describiste, tener esta disposición es peor para ellas. Esto es un mal

efecto en los términos de PI. Si la creencia en PI tiene este efecto, PI falla en sus propios términos».

Para responder a esta objeción, necesitamos saber si estas personas pueden cambiar su disposición. Supongamos, en primer lugar, que no pueden. ¿Por qué sería esto verdadero? Si no pueden cambiar su disposición, y tienen esta disposición *porque* creen en PI, la explicación tiene que ser que no pueden determinarse a sí mismas a estar dispuestas a hacer lo que creen que es irracional. Podrían cambiar su disposición sólo si creyesen en alguna otra teoría de la racionalidad. PI les diría entonces que, si pueden, se determinen a sí mismas a creer en esta otra teoría. Discuto esta posibilidad en las Secciones 6 a 8. Como defenderé, aun si esto es verdadero PI no fallaría en sus propios términos.

Supongamos a continuación que estas personas pueden cambiar su disposición sin cambiar sus creencias sobre la racionalidad. Si estas personas no son nunca abnegadas, esto será peor para ellas que si tuvieran alguna otra disposición. PI dice a estas personas que se determinen a sí mismas a tener esta otra disposición. La objeción dada arriba fracasa con toda claridad. Puede ser verdadero que estas personas nunca sean abnegadas porque creen en PI. Pero PI afirma que es irracional para estas personas permitirse seguir no siendo nunca abnegadas. Si siguen no siendo nunca abnegadas, no puede afirmarse que esto sea meramente «el resultado de su creencia en PI». Es el resultado de su fracaso a la hora de hacer lo que podrían hacer y lo que PI les dice que hagan. Este resultado es peor para ellas, lo que es un mal efecto en los términos de PI. Pero un mal efecto que resulta de *desobedecer* PI no puede suministrar una objeción contra PI. Si mi médico me dice que me vaya a vivir a un clima más saludable, no sería objeto de ninguna crítica si, a consecuencia de mi negativa a marcharme, yo muriera.

Hay una tercera posibilidad. Estas personas pueden ser incapaces de cambiar o sus disposiciones o sus creencias sobre la racionalidad. Su creencia en PI es mala para ellas, lo que es un mal efecto en los términos de PI. ¿Es esto una objeción contra PI? Será más fácil contestar a esta pregunta cuando hayamos discutido otras teorías. Mi respuesta está en la Sección 18.

5. ¿Podría ser racional determinarse a sí mismo a actuar irracionalmente?

Me vuelvo ahora a una nueva cuestión. Una teoría puede ser inaceptable aunque no falle en sus propios términos. Es cierto de muchas personas que sería peor para ellas no ser nunca abnegadas. ¿Nos da esto fundamentos independientes para rechazar PI?

Según PI, sería racional para cada una de estas personas determinarse a tener, o a mantener, uno de los mejores conjuntos de motivos posibles, en términos del propio interés. Cuáles sean estos conjuntos es en parte una cuestión fáctica. Y los detalles de la respuesta serían diferentes para diferentes personas en diferentes circunstancias. Pero de todas estas personas sabemos lo siguiente. Puesto que sería peor para ellas no ser nunca abnegadas, sería mejor para ellas ser a veces abnegadas. Sería mejor para ellas estar a veces dispuestas a hacer lo que creyesen que iba a ser peor para ellas. PI afirma que actuar de esta manera es irracional. Si una tal persona cree en PI, la teoría le dice que se determine a estar dispuesto a actuar de un modo que PI afirma que es irracional. ¿Es esta una implicación perjudicial? ¿Nos da alguna razón para rechazar PI?

Consideremos

La Respuesta de Schelling al Robo a Mano Armada. Un hombre irrumpe en mi casa. Me oye llamar a la policía, pero, como la ciudad más próxima está lejos, la policía no podrá llegar hasta dentro de por lo menos quince minutos. El ladrón me manda abrir la caja fuerte en la que guardo mis joyas de oro, amenazándome con que si no se las doy en cinco minutos empezará a matar a tiros a mis hijos, uno por uno.

¿Cuál sería para mí la conducta más racional? Necesito la respuesta pronto. Por una parte me doy cuenta de que no sería racional darle al ladrón el oro. Porque sabe que si se limita a llevárselo, o mis hijos o yo mismo le podríamos decir a la policía la marca y el número de la matrícula del coche en el que se iría. Así que corremos un gran riesgo de que, en caso de conseguir las joyas, nos mate a todos antes de huir.

Como sería irracional darle al ladrón las joyas, ¿debería ignorar entonces su amenaza? Pero esto también sería irracional. Corro un

efecto en los términos de PI. Si la creencia en PI tiene este efecto, PI falla en sus propios términos».

Para responder a esta objeción, necesitamos saber si estas personas pueden cambiar su disposición. Supongamos, en primer lugar, que no pueden. ¿Por qué sería esto verdadero? Si no pueden cambiar su disposición, y tienen esta disposición *porque* creen en PI, la explicación tiene que ser que no pueden determinarse a sí mismas a estar dispuestas a hacer lo que creen que es irracional. Podrían cambiar su disposición sólo si creyesen en alguna otra teoría de la racionalidad. PI les diría entonces que, si pueden, se determinen a sí mismas a creer en esta otra teoría. Discuto esta posibilidad en las Secciones 6 a 8. Como defenderé, aun si esto es verdadero PI no fallaría en sus propios términos.

Supongamos a continuación que estas personas pueden cambiar su disposición sin cambiar sus creencias sobre la racionalidad. Si estas personas no son nunca abnegadas, esto será peor para ellas que si tuvieran alguna otra disposición. PI dice a estas personas que se determinen a sí mismas a tener esta otra disposición. La objeción dada arriba fracasa con toda claridad. Puede ser verdadero que estas personas nunca sean abnegadas porque creen en PI. Pero PI afirma que es irracional para estas personas permitirse seguir no siendo nunca abnegadas. Si siguen no siendo nunca abnegadas, no puede afirmarse que esto sea meramente «el resultado de su creencia en PI». Es el resultado de su fracaso a la hora de hacer lo que podrían hacer y lo que PI les dice que hagan. Este resultado es peor para ellas, lo que es un mal efecto en los términos de PI. Pero un mal efecto que resulta de *desobedecer* PI no puede suministrar una objeción contra PI. Si mi médico me dice que me vaya a vivir a un clima más saludable, no sería objeto de ninguna crítica si, a consecuencia de mi negativa a marcharme, yo muriera.

Hay una tercera posibilidad. Estas personas pueden ser incapaces de cambiar o sus disposiciones o sus creencias sobre la racionalidad. Su creencia en PI es mala para ellas, lo que es un mal efecto en los términos de PI. ¿Es esto una objeción contra PI? Será más fácil contestar a esta pregunta cuando hayamos discutido otras teorías. Mi respuesta está en la Sección 18.

5. ¿PODRÍA SER RACIONAL DETERMINARSE A SÍ MISMO A ACTUAR IRRACIONALMENTE?

Me vuelvo ahora a una nueva cuestión. Una teoría puede ser inaceptable aunque no falle en sus propios términos. Es cierto de muchas personas que sería peor para ellas no ser nunca abnegadas. ¿Nos da esto fundamentos independientes para rechazar PI?

Según PI, sería racional para cada una de estas personas determinarse a tener, o a mantener, uno de los mejores conjuntos de motivos posibles, en términos del propio interés. Cuáles sean estos conjuntos es en parte una cuestión fáctica. Y los detalles de la respuesta serían diferentes para diferentes personas en diferentes circunstancias. Pero de todas estas personas sabemos lo siguiente. Puesto que sería peor para ellas no ser nunca abnegadas, sería mejor para ellas ser a veces abnegadas. Sería mejor para ellas estar a veces dispuestas a hacer lo que creyesen que iba a ser peor para ellas. PI afirma que actuar de esta manera es irracional. Si una tal persona cree en PI, la teoría le dice que se determine a estar dispuesto a actuar de un modo que PI afirma que es irracional. ¿Es esta una implicación perjudicial? ¿Nos da alguna razón para rechazar PI?

Consideremos

La Respuesta de Schelling al Robo a Mano Armada. Un hombre irrumpe en mi casa. Me oye llamar a la policía, pero, como la ciudad más próxima está lejos, la policía no podrá llegar hasta dentro de por lo menos quince minutos. El ladrón me manda abrir la caja fuerte en la que guardo mis joyas de oro, amenazándome con que si no se las doy en cinco minutos empezará a matar a tiros a mis hijos, uno por uno.

¿Cuál sería para mí la conducta más racional? Necesito la respuesta pronto. Por una parte me doy cuenta de que no sería racional darle al ladrón el oro. Porque sabe que si se limita a llevárselo, o mis hijos o yo mismo le podríamos decir a la policía la marca y el número de la matrícula del coche en el que se iría. Así que corremos un gran riesgo de que, en caso de conseguir las joyas, nos mate a todos antes de huir.

Como sería irracional darle al ladrón las joyas, ¿debería ignorar entonces su amenaza? Pero esto también sería irracional. Corro un

gran riesgo de que mate a uno de mis hijos para convencerme de que su amenaza de que como no le dé el oro los mata a todos va en serio.

¿Qué debería hacer yo? Es muy probable que, le dé o no le dé las joyas al ladrón, nos mate a todos. Mi situación es desesperada. Por fortuna recuerdo haber leído *La estrategia del conflicto* [3], de Schelling. También dispongo de una droga especial, por fortuna al alcance de la mano. Esta droga le hace a uno ser muy irracional durante un breve periodo de tiempo. Antes de que el hombre pueda detenerme, cojo la botella y bebo. En unos cuantos segundos, queda patente que me he vuelto loco. Haciendo eses por la habitación, le digo al ladrón: «¡Adelante! ¡Quiero a mis hijos, así que por favor mátelos!». El hombre intenta conseguir las joyas torturándome, pero grito: «¡Qué dolor tan terrible!, ¡siga, se lo ruego!».

Teniendo en cuenta mi estado, el hombre se ve ahora impotente. Y es que no puede hacer nada para inducirme a abrir la caja fuerte. Ni las amenazas ni la tortura pueden forzar concesiones en alguien que es tan irracional. Lo único que puede hacer es largarse, con la esperanza de escapar de la policía. Y, como yo me encuentro en este estado, es menos probable que crea que vaya a recordar el número de la matrícula de su coche. Así que no tiene ninguna razón para asesinarme.

Mientras esté en este estado, actuaré de forma irracional. Corro el riesgo de que, antes de la llegada de la policía, me dañe a mí mismo o lastime a mis hijos. Pero, como no tengo armas, el riesgo es pequeño. Y convertirme en irracional es el mejor modo de reducir el gran riesgo de que el ladrón nos mate a todos.

Según cualquier teoría plausible de la racionalidad, sería racional para mí, en este caso, determinarme a mí mismo a volverme irracional durante un lapso de tiempo [4]. Esto responde la pregunta que formulé arriba. PI podría decirnos que nos determináramos a estar dispuestos a actuar de maneras que PI afirma son irracionales. Esto no es objeción contra PI. Como el caso que acabamos de exponer demuestra, una teoría aceptable de la racionalidad *puede* decirnos

[3] Véase Schelling (1).
[4] Véase Nagel (1), capt. V.

que nos determinemos a hacer lo que, en sus propios términos, es irracional.

Consideremos a continuación una afirmación general que en ocasiones se hace:

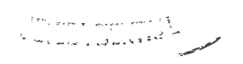
> (G1) Si hay algún motivo tal que fuese tanto (a) racional para alguien determinarse a tenerlo, como (b) irracional para él determinarse a perderlo, entonces (c) no puede ser irracional para esta persona actuar sobre la base de este motivo

En el caso que acabamos de describir, mientras el ladrón se halla todavía en mi casa, sería irracional para mi determinarme a dejar de ser irracional. Durante este periodo, tengo un conjunto de motivos de los que tanto (a) como (b) son verdaderos. Pero (c) es falso. Durante este periodo, mis actos son irracionales. Por tanto deberíamos rechazar (G1). Podemos afirmar en vez de esto que, puesto que era racional para mí determinarme a ser así, este es un caso de irracionalidad *racional*.

6. Cómo PI implica que no podemos evitar actuar irracionalmente

Recordemos a Kate, que aceptaba la Teoría Hedonista del propio interés. Nosotros podemos aceptar alguna otra teoría. Pero según estas otras teorías podrían darse casos que, en los aspectos relevantes, fuesen como el de Kate. Y las afirmaciones que siguen podrían ser reformuladas para cubrir estos casos.

Es lo mejor para Kate que su deseo más fuerte sea que sus libros sean los mejores posibles. Pero, como esto es cierto, a menudo trabaja muy duramente, quedándose agotada y deprimida durante una larga temporada. Como Kate es hedonista, cree que al actuar así está haciendo lo que es peor para ella. Como también acepta PI, Kate cree que, en estos casos, está actuando irracionalmente. Además, estos actos irracionales son completamente voluntarios. Actúa como lo hace porque, aunque se cuida de sus propios intereses, este no es su deseo más fuerte. Tiene un deseo más fuerte aún, que sus



libros sean los mejores posibles. Y sería peor para ella que este deseo se debilitara. Está actuando a partir de un conjunto de motivos que, de acuerdo con PI, sería para ella irracional determinarse a perder.

Podría decirse que, puesto que Kate está actuando a partir de tales motivos, no puede estar actuando irracionalmente. Pero esta afirmación asume (G1), la afirmación cuya falsedad mostró el caso que llamé la Respuesta de Schelling al Robo a Mano Armada.

Si compartimos la creencia de Kate de que está actuando irracionalmente de un modo absolutamente voluntario, podríamos afirmar que *ella* es irracional. Pero Kate puede negarlo. Como cree en PI, puede afirmar: «Cuando hago lo que creo que será peor para mí, mi *acto* es irracional. Pero, como estoy actuando a partir de un conjunto de motivos que sería irracional para mí determinarme a perder, *yo no* soy irracional. Más precisamente, soy *racionalmente irracional*».

Y puede añadir: «Al actuar según mi deseo de hacer mis libros mejores, estoy haciendo lo que será peor para mí. Este es un mal efecto, en los términos del propio interés. Pero forma parte de un conjunto de efectos que es uno de los mejores conjuntos posibles. Aunque a veces yo sufra, como este es mi más fuerte deseo también me beneficio. Y los beneficios son más grandes que las pérdidas. Que yo a veces actúe irracionalmente, haciendo lo que sé que será peor para mí, es el precio que tengo que pagar si voy a conseguir estos beneficios mayores. Este es un precio que vale la pena pagar».

Puede objetarse: «Tú no *tienes* que pagar este precio. *Podrías* trabajar menos duramente. Podrías hacer lo que fuese mejor para ti. No estás obligada a hacer lo que crees irracional».

Ella podría contestar: «Cierto. *Podría* trabajar menos duramente. Pero yo sólo lo *haría* si mi deseo de mejorar mis libros fuese mucho más débil. Y esto sería, vistas las cosas en su conjunto, peor para mí. Mi trabajo se volvería aburrido. ¿Cómo podría yo ocasionar que no fuese a elegir libremente hacer en el futuro, en casos así, lo que creo irracional? Yo podría ocasionarlo sólo cambiando mis deseos de un modo que fuese peor para mí. Este es el sentido en que no puedo tener los mayores beneficios sin pagar los menores precios. No puedo tener los deseos que son los mejores para mí sin

elegir libremente de vez en cuando actuar de maneras que son peores para mí. Por esta razón, cuando actúo irracionalmente de estas formas, no necesito considerarme *a mí misma* irracional».

Esta réplica da por buena una determinada concepción de los actos voluntarios: *el determinismo psicológico*. Según esta visión, nuestros actos son siempre causados por nuestros deseos, creencias y otras disposiciones. Dados nuestros deseos y disposiciones reales, no es causalmente posible que actuemos de diferente manera. Podría objetarse: «Si no es causalmente posible que Kate actúe de forma diferente, ella no debería creer que, para actuar racionalmente, *debe de* actuar de forma diferente. Sólo *debemos* hacer lo que *podemos* hacer».

Surgirá más tarde una objeción similar cuando discuta lo que moralmente debemos hacer. Ahorrará palabras el que Kate responda a las dos objeciones. Ella puede decir: «En la doctrina de que *deber* implica *poder*, el sentido de «poder» es compatible con el determinismo psicológico. Cuando mi acto es irracional o malo, yo debería de haber actuado de alguna otra manera. Según esta doctrina, yo debería haber actuado de otro modo sólo si yo hubiera podido hacerlo así. Si yo *no* hubiera podido actuar de otro modo, no puede afirmarse que esto es lo que yo debería haber hecho. La afirmación (1) de que yo no podía haber actuado de otro modo no es la afirmación (2) de que actuar de este modo habría sido imposible, dados mis deseos y disposiciones reales. La afirmación es más bien (3) que actuar de otro modo habría sido imposible, aunque mis deseos y disposiciones hubiesen sido diferentes. Actuar de otro modo habría sido imposible, fueran cuales fueran mis deseos y disposiciones. Si la afirmación (1) fuese la afirmación (2), los deterministas tendrían que concluir que nunca es posible para nadie actual mal o irracionalmente. Pero pueden rechazar esta conclusión de forma justificada. Pueden insistir en que la afirmación (1) es la afirmación (3)».

Kate podría añadir: «No estoy afirmando que la *voluntad libre* sea compatible con el determinismo. El sentido de «poder» que se requiere para la voluntad libre puede ser diferente del sentido de «poder» en la doctrina de que deber implica poder. La mayoría de los determi-

nistas que creen que la voluntad libre *no* es compatible con el determinismo mantienen que estos sentidos son diferentes. Por eso, aunque estos deterministas no crean que nadie merece castigo, continúan creyendo que es posible actuar mal o irracionalmente».

Kate puede equivocarse al asumir el determinismo psicológico. Antes afirmé que nuestras creencias sobre la racionalidad pueden afectar a nuestros actos porque podemos querer actuar racionalmente. Puede objetarse:

> Esto describe mal el modo en que nuestras creencias afectan a nuestros actos. Nosotros no *explicamos* por qué alguien ha actuado racionalmente citando su deseo de hacerlo. Cuandoquiera que alguien actúa racionalmente puede ser trivialmente verdadero que lo quería hacer así. Pero actuó como lo hizo porque tenía una creencia, no una creencia *y* un deseo. Actuó como lo hizo simplemente porque creía tener una razón para hacerlo. Y a menudo es causalmente posible para él actuar racionalmente tenga los deseos y las disposiciones que tenga [4a]*.

Hay que notar que este objetor no puede afirmar que sea *siempre* posible para alguien actuar racionalmente, sean los que sean sus deseos y disposiciones. Aunque niegue el determinismo, este objetor no puede afirmar que *no* haya conexión alguna entre nuestros actos y nuestras disposiciones.

Este objetor tiene que admitir también que nuestros deseos y disposiciones pueden *dificultar* que hagamos lo que creemos racional. Supongamos que estoy sufriendo una sed intensa, y que me dan un vaso de agua helada. Y supongamos que yo creo tener una razón para beber este agua despacio, puesto que esto incrementaría mi disfrute. Tengo también una razón para no derramar este agua. Es mucho más fácil actuar sobre la base de esta segunda razón de lo que lo es, dada mi intensa sed, beber esta agua lentamente.

[4a] Suponiendo, como podría ser verdadero, que yo no pudiera hacerme meramente aparecer a mí mismo como siendo irracional.

* Por varias razones, hemos decidido ser fieles al modo de citar del original, sin duda poco convencional: Parfit a menudo da saltos de uno o más números o, como en este caso, le coloca un índice a la cita. (N. del t.)

Si las afirmaciones del objetor fuesen verdaderas, la réplica de Kate tendría que ser revisada. Ella podría decir: «Sería peor para mí que mi deseo más fuerte fuese evitar hacer lo que yo creo que es irracional. Es mejor para mí que mi deseo más fuerte sea que mis libros sean los mejores posibles. Puesto que este es mi más fuerte deseo, a veces hago lo que creo irracional. Actúo así porque mi deseo de mejorar mis libros es mucho más fuerte que mi deseo de no actuar irracionalmente. Tú afirmas que a menudo podría evitar actuar así. Por un esfuerzo de voluntad, a menudo podría yo evitar hacer lo que tengo más ganas de hacer. Si yo pudiera evitar actuar así, no podría afirmar que no soy en ningún sentido irracional. Pero, supuesta la fuerza de mi deseo de mejorar mis libros, sería *muy difícil* para mí evitar actuar así. Y sería para mí irracional cambiar mis deseos de forma que fuese más fácil para mí evitar actuar así. Supuestos estos hechos, sólo soy irracional en un sentido muy débil».

Kate podría añadir: «No es posible que se den estas *dos* cosas, que yo tenga uno de los mejores conjuntos posibles de motivos, en términos del propio interés, *y* que yo nunca haga lo que considero irracional. Esto no es posible en el sentido relevante: no es posible *sean los que sean* mis deseos y disposiciones. Si no fuera nunca abnegada, mis actos corrientes nunca serían irracionales. Pero yo habría actuado irracionalmente al determinarme a mí misma a no ser nunca abnegada, o al permitirme a mí misma continuar con esta disposición. Si en cambio me determino a tener uno de los mejores conjuntos de motivos, algunas veces haré lo que creo irracional. Si no tengo la *disposición* de alguien que no es nunca abnegado, no es posible que yo *siempre actúe* como alguien con esta disposición. Puesto que esto no es posible, y sería para mí irracional determinarme a no ser nunca abnegada, no se me puede criticar porque algunas veces haga lo que creo irracional».

Ahora se puede decir que, tal y como la describió Kate, PI carece de uno de los rasgos esenciales de cualquier teoría. Se puede objetar: «Ninguna teoría puede exigir lo que es imposible. Ya que Kate no siempre puede evitar hacer lo que PI afirma que es irracional, no siempre puede hacer lo que PI afirma que debiera

hacer. Por tanto deberíamos rechazar PI. Como antes, *deber* implica *poder*».

Aunque neguemos el determinismo, esta objeción todavía es aplicable. Como he afirmado, tenemos que admitir que, ya que Kate no tiene la disposición de alguien que no es nunca abnegado, ella no *siempre* puede actuar como una persona así.

¿Es una buena objeción contra PI el que Kate no siempre pueda evitar hacer lo que PI afirma que es irracional? Recordemos la Respuesta de Schelling al Robo a Mano Armada. En este ejemplo, según cualquier teoría plausible de la racionalidad, sería irracional para mí no convertirme en alguien muy irracional. Pero si me convierto a mí mismo en alguien muy irracional, no podré evitar actuar irracionalmente. Según ambas alternativas, al menos uno de mis actos sería irracional. Por tanto es verdadero que, en este caso, no puedo evitar actuar irracionalmente. Ya que pueden darse casos tales, una teoría aceptable puede implicar que no podemos evitar actuar irracionalmente. No supone ninguna objeción contra PI el tener esta implicación.

Podemos pensar que estas afirmaciones no anulan completamente esta objeción. Una objeción similar surgirá más adelante contra ciertas teorías morales. Para ahorrar palabras, discutiré estas objeciones juntas en la Sección 15.

Resumiré ahora mis otras conclusiones. En el caso de muchas personas, quizás de la mayoría, la teoría del Propio Interés es indirectamente contraproducente. Es verdadero, de cada una de estas personas, que sería peor para ellas no ser nunca abnegado —no estar nunca dispuesto a hacer lo que creemos que va a ser peor para nosotros—. Sería mejor para ellas tener algún otro conjunto de motivos. He afirmado que tales casos no proporcionan una objeción contra PI . Puesto que PI no les dice a estas personas que no sean nunca abnegadas, sino que lo que les dice es que, si pueden, lo sean, PI no falla en sus propios términos. Tampoco proporcionan estos casos una objeción independiente contra PI.

Aunque no refutan PI, para aquellos que aceptan PI estos casos son de gran importancia. En estos casos, PI tiene que cubrir no sólo actos corrientes sino también los actos que introducen cambios en

nuestros motivos. Según PI, sería racional determinarnos a tener, o a mantener, uno de los mejores conjuntos de motivos posibles, en términos de propio interés. Si creemos que podríamos actuar de alguna de estas maneras, sería irracional no hacerlo así. En el caso de la mayoría de las personas, cualquiera de los mejores conjuntos posibles determinaría que estas personas a veces hicieran, de modo completamente voluntario, lo que saben que será peor para ellas. Si estas personas creen en PI, creerán que estos actos son irracionales. Pero no es necesario que crean que *ellas mismas* son irracionales. Esto es porque, de acuerdo con PI, sería irracional para ellas cambiar sus motivos de modo que dejaran de actuar irracionalmente de esta manera. En parte lamentarán las *consecuencias* de estos actos irracionales. Pero podrán considerar con complacencia la *irracionalidad* de estos actos. Esto es irracionalidad *racional*.

Puede objetarse a estas afirmaciones que asumen falsamente el determinismo psicológico. En ocasiones puede ser posible para estas personas hacer lo que creen racional, sean los que sean sus deseos y disposiciones. Si esta objeción es correcta, estas afirmaciones necesitan revisión. Cuando estas personas hacen lo que creen irracional, no pueden afirmar que no son irracionales en ningún sentido. Pero lo que sí pueden afirmar es que, supuestos sus motivos reales, sería muy difícil para ellas evitar actuar de este modo. Y sería irracional para ellas, según su teoría, cambiar sus motivos de manera que fuese más fácil evitar actuar de este modo. Por tanto pueden afirmar que son irracionales sólo en un sentido muy débil. Habiendo explicado una vez cómo podrían ser revisadas estas afirmaciones, no mencionaré esta objeción cuandoquiera que sea relevante en lo que sigue. Sería fácil hacerle las revisiones necesarias a cualquier afirmación similar.

7. Un argumento para rechazar PI cuando entra en conflicto con la moralidad

Se ha sostenido que la teoría del Propio Interés podría decirnos que creyéramos, no en ella misma, sino en alguna otra teoría. Sin duda

esto es posible. Según PI, sería racional para nosotros determinarnos a creer alguna otra teoría, si esto fuese mejor para nosotros.

Ya he mencionado un modo en que esto podría ocurrir. No sería posible para nosotros hacer lo que creemos irracional. PI nos diría entonces, en los casos que he estado discutiendo, que intentáramos creer en una teoría diferente. Hay también otros modos en que esto podría ocurrir. Volvamos, para poner un ejemplo, al cumplimiento de nuestras promesas.

Tiene gran importancia práctica cierto tipo de acuerdos mutuos. En estos acuerdos, cada persona de determinado grupo hace una promesa condicional. Cada persona promete actuar de cierto modo, contando con que todas las demás prometan actuar de cierto modo. Puede ser verdadero tanto que (1) será mejor para cada una de estas personas el que todas más bien que ninguna cumplan sus promesas, y que (2) sea lo que sea lo que las demás hagan, será peor para cada persona cumplir su promesa. Lo que cada persona pierde si cumple su promesa es menos de lo que gana si todas las demás cumplen sus promesas. Así es como (1) y (2) son los dos verdaderos. Tales acuerdos son *mutuamente ventajosos, aunque requieren abnegación.*

Si se sabe que nunca soy abnegado, me excluirán de tales acuerdos. Los demás sabrán que no se puede confiar en que yo cumpla mis promesas. Se ha afirmado que, puesto que esto es verdadero, sería mejor para mí si yo dejara de no ser nunca abnegado y me hiciera digno de confianza [5].

Esta afirmación pasa por alto una posibilidad. Lo mejor para mí puede ser *parecer* digno de confianza pero seguir sin ser nunca abnegado. Puesto que parezco digno de confianza, los demás me admitirán en estos acuerdos mutuamente ventajosos. Como en realidad nunca soy abnegado, me beneficiaré rompiendo mis promesas siempre que sea mejor para mí. Puesto que es mejor para mí parecer digno de confianza, a menudo será mejor para mí cumplir mi promesa para preservar esta apariencia. Pero habrá algunas promesas que podré romper en secreto. Y la ganancia que obtendré de romper algunas promesas puede pesar más que mi pérdida al dejar de parecer digno de confianza.

[5] Esto se afirma en Gauthier (3) y (4).

Pero supongamos que soy transparente, incapaz de mentir convincentemente. Esto le sucede a mucha gente. Y podría suceder mucho más si desarrolláramos pruebas baratas y fiables de detección de mentiras. Asumamos que esto ha ocurrido, de modo que todos somos transparentes —incapaces de engañar a los demás—. Como nosotros somos transparentes hasta cierto punto, mis conclusiones pueden ser de aplicación a nuestra situación real. Pero simplificará el argumento asumir que todo engaño directo se ha hecho imposible. Vale la pena ver lo que semejante argumento podría revelar. Por eso deberíamos facilitar el argumento, dando por válida esta asunción.

Si fuéramos todos transparentes, sería mejor para cada uno de nosotros llegar a ser digno de confianza: dispuestos fidedignamente a cumplir nuestras promesas, aun cuando creamos que hacerlo así será peor para nosotros. Por tanto, según PI, sería racional para cada uno de nosotros hacerse digno de confianza.

Asumamos además que, para hacernos dignos de confianza, tuviéramos que cambiar nuestras creencias sobre la racionalidad. Tendríamos que hacernos creer que es racional para cada uno de nosotros cumplir sus promesas, aun cuando sepamos que esto sería peor para nosotros. Después describiré dos modos en que esta asunción podría ser verdadera.

Es difícil cambiar nuestras creencias cuando nuestra razón para hacerlo así es simplemente que este cambio favorecerá nuestros intereses. Tendríamos que utilizar alguna forma de auto-engaño. Supongamos, por ejemplo, que me entero de que estoy mortalmente enfermo. Como quiero creer que estoy sano, pago a un hipnotizador para que me dé esta creencia. No podría mantener esta creencia si recordara cómo la había adquirido. Porque si lo recordara, sabría que la creencia era falsa. Lo mismo podría ser verdadero de nuestras creencias sobre la racionalidad. Si pagamos a hipnotizadores para que cambien estas creencias porque tal cosa es mejor para nosotros, los hipnotizadores tendrían que hacernos olvidar por qué tenemos nuestras nuevas creencias.

Según las asunciones hechas arriba, PI nos diría que cambiáramos nuestras creencias. PI nos diría que la creyésemos no a ella misma, sino a una forma revisada de PI. Según esta teoría revisada, es irracional para nosotros hacer lo que pensamos que es peor para nosotros, *excepto cuando cumplimos una promesa.*

Si PI nos dijera que creyésemos en esta teoría revisada, ¿sería esto una objeción a PI? ¿Mostraría que *es* racional cumplir tales promesas? Tenemos que concentrarnos sin dudarlo en esta pregunta. Puede que tengamos razón al creer que es racional cumplir nuestras promesas aun cuando sepamos que esto va a ser peor para nosotros. Lo que pregunto es: «¿Tendría apoyo esta creencia si PI misma nos dijese que nos hiciésemos tener esta creencia?».

Algunos contestan Sí. Mantienen que si PI nos dice que nos determinemos a tener esta creencia, esto revela que esta creencia está justificada. Y aplican este argumento a muchos otros tipos de acto de los que piensan que vienen moralmente exigidos, como cumplir las promesas. Si este argumento tuviera éxito, ello tendría una gran importancia. Porque revelaría que en muchos tipos de caso es racional actuar moralmente, aun cuando creamos que esto va a ser peor para nosotros. Se habría demostrado que las razones morales son más poderosas que las razones que proporciona el propio interés. Muchos escritores han intentado justificar esta conclusión sin éxito. Si esta conclusión pudiera justificarse del modo que acabo de mencionar, se resolvería lo que Sidgwick llamó «el problema más profundo de la Ética» [6].

8. Por qué falla este argumento

Hay una objeción simple contra este argumento. El argumento apela al hecho de que PI nos diría que nos hiciéramos a nosotros mismos creer que es racional cumplir nuestras promesas aun cuando supiéramos que esto iba a ser peor para nosotros. Llamemos a esta creencia *B*. B es incompatible con PI, puesto que

[6] Sidgwick (1), p. 386, nota al pie 4.

PI afirma que es irracional cumplir tales promesas. O bien PI es la teoría verdadera de la racionalidad, o bien no lo es. Si PI es verdadera, B tiene que ser falsa, ya que es incompatible con PI. Si PI no es verdadera, B podría ser verdadera, pero PI no puede apoyar a B, puesto que una teoría que no es verdadera no puede apoyar ninguna conclusión. Brevemente: si PI es verdadera, B tiene que ser falsa, y si PI no es verdadera, no puede servir de apoyo a B. B o bien es falsa o bien no está apoyada. De modo que, aun cuando PI nos diga que intentemos creer en B, este hecho no puede servir de apoyo a B.

Podemos pensar que una teoría de la racionalidad no puede ser verdadera sino que puede, como máximo, ser la mejor teoría, o la mejor justificada. La objeción que acabamos de examinar podría ser reformulada en estos términos. Hay dos posibilidades. Si PI es la mejor teoría, deberíamos rechazar B, puesto que es incompatible con PI. Si PI no es la mejor teoría, deberíamos rechazar PI. Y B no puede ser apoyada por una teoría que debemos rechazar. Ninguna de estas posibilidades puede dar ningún apoyo a B [7].

Esta objeción me parece fuerte. Pero conozco a algunos a quienes no les convence. Por eso pondré dos objeciones más. Estas objeciones apoyarán también algunas conclusiones más amplias.

En primer lugar distinguiré las amenazas de las advertencias. Cuando diga que haré X como no hagas Y, denominaré a esto una *advertencia* si mi hacer X sería peor para ti pero no para mí, y una *amenaza* si mi hacer X sería peor para los dos. Llamadme un *ejecutor de amenazas* si yo siempre cumplo mis amenazas.

Supongamos que alguien, además de ser un ejecutor de amenazas, no es nunca abnegado. Tal persona cumpliría sus amenazas aunque supiera que esto iba a ser peor para ella. Pero nunca *haría* amenazas si creyera que hacerlo iba a ser peor para ella. Esto es así porque, además de ser una ejecutora de amenazas, esta persona nunca es abnegada. Nunca hace lo que cree que va a ser peor para ella, *excepto cuando se trata de ejecutar alguna amenaza*. Esta excepción no cubre *hacer* amenazas.

[7] Estas objeciones me las indicó S. Kagan.

Supongamos que todos nosotros somos transparentes y nunca somos abnegados. Si esto fuera verdadero, sería mejor para mí que me convirtiese en un ejecutor de amenazas, y que después anunciara a todos los demás este cambio en mis disposiciones. Como soy transparente, todos creerían mis amenazas. Y que se creyesen mis amenazas sería de mucha utilidad. Algunas de mis amenazas podrían ser defensivas, intentando protegerme de agresiones externas. Yo podría limitarme a las amenazas defensivas. Pero sería tentador usar mi conocida disposición de otros modos. Supongamos que los beneficios de alguna cooperación se reparten entre nosotros. Y supongamos que, sin mi cooperación, no habría más beneficios. Yo podría decir que, a no ser que obtenga la participación más grande, no cooperaré. Si los demás saben que yo soy un ejecutor de amenazas, y ellos no son nunca abnegados, me darán la participación más grande. No hacerlo así sería peor para ellos.

Otros ejecutores de amenazas podrían actuar peor. Podrían reducirnos a la esclavitud. Podrían amenazarnos con que, como no nos convirtamos en sus esclavos, provocarán nuestra destrucción mutua. Nosotros sabríamos que estas personas realizarían sus amenazas. Por tanto sabríamos que la única manera de evitar la destrucción sería convirtiéndonos en sus esclavos.

La respuesta a los ejecutores de amenazas, si todos somos transparentes, no es otra que *ignorar la amenaza*. Una persona que responde así ignora la amenaza, aun cuando sabe que hacerlo será peor para ella. Un ejecutor de amenazas no amenazaría a uno que ignora las amenazas y que es transparente. Porque sabría que si lo hiciera, su amenaza sería ignorada, y él ejecutaría su amenaza, lo que iba a ser peor para él.

Si todos fuéramos transparentes, y nunca abnegados, ¿qué cambios en nuestras disposiciones iban a ser mejores para cada uno de nosotros? Respondo esta pregunta en el Apéndice A, ya que algunas partes de la respuesta no son relevantes en relación con la pregunta que estoy discutiendo ahora. Lo que es relevante es esto: si todos fuéramos transparentes, probablemente sería mejor para cada uno de nosotros convertirse en alguien digno de confianza que ignora las amenazas. Estos dos cambios implicarían ciertos riesgos; pero estos

riesgos serían compensados con mucho por los beneficios probables. ¿Qué beneficios reportaría hacerse digno de confianza? Que no seríamos excluidos de esos acuerdos mutuamente ventajosos que exigen abnegación. ¿Qué beneficios reportaría ignorar las amenazas? Que evitaríamos ser esclavos de los ejecutores de amenazas.

Podemos asumir a continuación que no podríamos ignorar las amenazas y ser dignos de confianza a no ser que cambiáramos nuestras creencias sobre la racionalidad. Los que son dignos de confianza cumplen sus promesas aun cuando sepan que esto va a ser peor para ellos. Podemos asumir que no podríamos ponernos en disposición de actuar así a no ser que creyéramos que *es* racional cumplir tales promesas. Y podemos asumir que, a no ser que supieran que tenemos esta creencia, los demás no iban a confiar en que cumpliéramos tales promesas. Partiendo de estos supuestos, PI nos dice que nos hagamos tener esta creencia. Observaciones similares se aplican a convertirse en los que ignoran las amenazas. Porque podemos asumir que no podríamos convertirnos en los que ignoran las amenazas a no ser que creyéramos que es siempre racional ignorar la amenazas. Y podemos asumir que, a no ser que tengamos esta creencia, los demás no estarían convencidos de que nosotros formamos parte de los que ignoran las amenazas. Partiendo de estos supuestos, PI nos dice que nos hagamos tener esta creencia. Y estas conclusiones pueden combinarse. PI nos dice que nos hagamos creer que es siempre irracional hacer lo que creemos que va a ser peor para nosotros, *excepto cuando cumplimos promesas o ignoramos amenazas*.

¿Apoya este hecho estas creencias? Según PI, sería racional para cada uno de nosotros hacerse creer que ignorar las amenazas es racional, aunque sepa que esto será peor para él. ¿Muestra esto que esta creencia es correcta? ¿Muestra que *es* racional ignorar tales amenazas?

Ayudará disponer de un ejemplo. Consideremos

> *Mi esclavitud*. Tú y yo compartimos una isla desierta. Ambos somos transparentes, y nunca abnegados. De repente introduces un cambio en tus disposiciones, y te conviertes en un ejecutor de amenazas. Y tienes una bomba, que podría hacer que la isla explotara.

> Amenazándome a menudo con hacer explotar esta bomba, me obligas a trabajar duro para ti. El único límite que tienes que respetar es dejar que mi vida siga valiendo la pena. Si mi vida dejara de valer la pena, ya no sería mejor para mí ceder a tus amenazas.

¿Qué puedo hacer para poner punto final a mi esclavitud? No serviría de nada matarte, puesto que tu bomba explotaría automáticamente si no marcaras regularmente cierto número secreto. Pero supongamos que yo pudiera convertirme de manera transparente en alguien que ignora las amenazas. Estúpidamente, no me has amenazado con que ignorarías este cambio en mis disposiciones. Así que este cambio pondría fin a mi esclavitud.

¿Sería racional para mí dar lugar a este cambio? Existe el riesgo de que tú podrías amenazarme con algo nuevo. Pero como hacer esto sería sin duda peor para ti, el riesgo sería pequeño. Y aceptando este pequeño riesgo, casi seguro que yo obtendría un beneficio muy grande. Casi seguro que pondría fin a mi esclavitud. Dado lo miserable de mi esclavitud, sería racional para mí, de acuerdo con PI, hacer por convertirme en alguien que ignora las amenazas. Y, dadas nuestras otras asunciones, sería racional para mí hacerme creer que es siempre racional ignorar las amenazas. Aunque no puedo estar completamente seguro de que esto vaya a ser mejor para mí, el beneficio, grande y casi seguro, superaría con mucho el pequeño riesgo. (Del mismo modo, nunca sería completamente seguro que fuese mejor para alguien ser digno de confianza. Aquí también todo lo que podría ocurrir es que los beneficios probables superaran a los riesgos.)

Concedamos que ahora he hecho estos cambios. Me he convertido de un modo transparente en alguien que ignora las amenazas, y me he determinado a mí mismo a creer que es siempre racional ignorar las amenazas. De acuerdo con PI, era racional para mí hacerme tener esta creencia. ¿Muestra esto que la creencia es correcta?

Continuemos la historia

> *Cómo termino con mi esclavitud.* Ambos tenemos mala suerte. Por un momento olvidas que me he convertido en alguien que ignora las amenazas. Para obtener algún fin sin importancia —por ejemplo, el

> coco que acabo de recoger— me reiteras tu amenaza de siempre. Me dices que como no te dé el coco harás que saltemos por los aires. Sé que si me niego va a ser seguramente peor para mí. Sé que con seguridad eres un ejecutor de amenazas, alguien que cumplirá sus amenazas aunque sepa que será peor para él. Pero, igual que tú, yo ahora creo en la teoría del puro Propio Interés. Ahora creo que es racional ignorar las amenazas, aunque sé que esto será peor para mí. Actúo sobre la base de esta creencia. Como preví, nos haces saltar por los aires.

¿Es mi acto racional? No lo es. Como antes, podríamos conceder que, puesto que yo actúo sobre una creencia que era racional para mí adquirir, *yo* no soy irracional. Más precisamente, yo soy *racionalmente* irracional. Pero lo que hago no es racional. Es irracional ignorar una amenaza cuando sé que, si lo hago, esto será desastroso para mí y mejor para nadie. PI me dijo aquí que era racional hacerme creer que es racional ignorar las amenazas, aun cuando sepa que esto va a ser peor para mí. Pero esto no demuestra que esta creencia sea correcta. No demuestra que, en semejante caso, *sea* racional ignorar las amenazas.

Podemos sacar una conclusión más amplia. Este caso nos demuestra que debemos rechazar

> (G2) Si es racional para alguien hacerse creer que es racional para él actuar de cierto modo, *es* racional para él actuar de cierto modo.

Volvamos ahora a B, la creencia de que es racional cumplir nuestras promesas, aunque sepamos que esto va a ser peor para nosotros. Según la asunción hecha arriba, PI implica que es racional para nosotros determinarnos a creer B. Algunos afirman que este hecho apoya B, demostrando que *es* racional cumplir tales promesas. Pero esta afirmación parece asumir (G2), que acabamos de rechazar.

Hay otra objeción a lo que estas personas afirman. Aunque PI nos diga que intentemos creer B, PI implica que B es falso. De modo que si B es verdadero, PI tiene que ser falso. Puesto que estas personas creen B, deben creer que PI es falso. Su afirmación asumiría entonces que

(G3) Si alguna teoría falsa de la racionalidad nos dice que nos hagamos tener una creencia particular, esto demuestra que esta creencia es verdadera.

Pero obviamente debemos rechazar (G3). Si alguna teoría falsa nos dijo que nos hiciéramos creer que la Tierra es plana, esto no mostraría que así es el caso.

PI nos dijo que intentásemos creer que es racional ignorar las amenazas, aun cuando sepamos que esto va a ser peor para nosotros. Pero, como mi ejemplo pone de manifiesto, esto no apoya esta creencia. Debemos por consiguiente hacer la misma afirmación por lo que respecta a cumplir las promesas. Puede haber *otros* fundamentos para creer que es racional cumplir nuestras promesas, aunque sepamos que hacerlo vaya a ser peor para nosotros. Pero esto no lo demostraría como racional el hecho de que la misma teoría del Propio Interés nos dijo que nos hiciésemos creer que era racional. Se ha mantenido que, apelando a tales hechos, podemos resolver un antiguo problema: podemos demostrar que, cuando se opone al propio interés, la moralidad proporciona las razones más fuertes para actuar. Este argumento falla. Lo máximo que podría mostrar es algo más pequeño. En un mundo en que todos fuéramos transparentes —incapaces de engañarnos unos a otros— podría ser racional engañarnos a nosotros mismos sobre la racionalidad [8].

[8] Podría por mi parte tratar de defender este argumento. Yo podría decir: «¿Por qué me dijo PI que creyera que es siempre racional ignorar las amenazas? Sólo porque tener esta creencia sería muy probablemente mejor para mí. Una vez que yo he tenido mala suerte, y tu has hecho tu amenaza, esta creencia deja de ser buena para mí. PI me diría que la perdiese. Por tanto este caso no demuestra lo que afirmé. Es irracional para mí ignorar tu amenaza. Pero cuando ignoro tu amenaza, PI ya no me dice que crea que esto es racional. Por eso podemos mantener la afirmación de que, si PI me dice que crea que un acto es racional, este acto es racional. Mi caso imaginario no aporta un contraejemplo. Cuando vamos de las amenazas a las promesas, podemos afirmar lo que negué. Como PI nos dice que creamos que es racional cumplir las promesas aunque sepamos que va a ser peor para nosotros, *es* racional cumplir tales promesas».

Esta defensa fracasa. ¿Por qué le dice PI a cada persona que tenga esta creencia? Porque si se sabe que no cree en esto, será excluida de la clase de acuerdos que

9. Cómo podría ser modesta PI

Si PI nos dijese que creyésemos en otra teoría, esto no serviría de apoyo a esta otra teoría. ¿Pero sería una objeción contra PI? Una vez más, PI no fallaría en sus propios términos. PI es una teoría de la racionalidad práctica, no de la racionalidad teórica. PI puede decirnos que nos las arreglemos para tener creencias falsas. Si fuese mejor para nosotros tener creencias falsas, tener creencias verdaderas, incluso acerca de la racionalidad, no sería parte del fin último que nos proporciona a nosotros PI.

Los argumentos dados arriba podrían fortalecerse y ampliarse. Esto sería más fácil si, como supuse, la tecnología de la detección de mentiras nos hiciera a todos completamente transparentes. Si nunca pudiéramos engañarnos los unos a los otros, podría haber un argumento que demostrara que, según PI, es racional para todos determinarse a no creer PI.

Supongamos que esto fuese verdadero. Supongamos que PI le dijese a todo el mundo que se determinase a creer en alguna otra teoría. PI sería entonces *modesto*. Si todos nosotros creyésemos en PI, pero pudiéramos también cambiar nuestras creencias, PI se quitaría a sí misma de la escena. Se convertiría en una teoría en la que nadie creería. Pero ser modesta no es lo mismo que ser contraproducente. Entre los fines de una teoría no está el que se crea en ella. Si personificáramos a las teorías y pretendiéramos que tuviesen fines, el fin de una teoría no sería ser creída, sino ser verdadera, o ser la mejor teoría. Que una teoría sea modesta no demuestra que no es la mejor teoría.

PI sería modesta, cuando, en caso de que creyéramos en PI, esto sería peor para nosotros. Pero PI no necesita decirnos que la crea-

describí, esos que, aunque mutuamente ventajosos, requieren abnegación. Supongamos que tengo esta creencia, y soy admitido a esos acuerdos. Aunque a menudo sería mejor para mí preservar mi reputación, habrá casos en que será peor para mí cumplir mis promesas. Cuando esto sea así, PI me dice que pierda mi creencia de que siempre es racional cumplir las promesas. No podemos decir que *es* racional para mí cumplir estas promesas porque PI me dice que crea en ello. Cuando cumplo estas promesas, PI ya no me dice que crea en ello.

mos. Cuando fuese mejor para nosotros creer en alguna otra teoría, PI nos diría que intentáramos creer en esta teoría. Si tuviésemos éxito en hacer lo que PI nos dijera que hiciéramos, esto sería mejor para nosotros una vez más. Aunque PI se quitase a sí misma de la escena, haciendo que nadie creyera en ella, aún no estaría fallando en sus propios términos. Aún sería verdadero que, porque todos hemos seguido PI —hemos hecho lo que PI nos dijo que hiciéramos— todos hemos hecho con ello que el resultado sea mejor para nosotros mismos.

Aunque PI no falle en sus propios términos, podría establecerse que una teoría aceptable no puede ser modesta. Niego esto. Puede parecer plausible por lo que, cuando lo examinamos, es una mala razón. Sería natural *querer* que la mejor teoría de la racionalidad no fuese modesta. Si la mejor teoría fuese modesta, y nos dijera que creyéramos en alguna otra teoría, la verdad sobre la racionalidad sería deprimentemente enrevesada. Es natural esperar que la verdad sea más simple: que la mejor teoría nos diga que creamos en ella. Pero ¿puede esto ser más que una esperanza? ¿Podemos dar por sentado que la verdad *tiene* que ser más simple? No podemos.

10. Cómo el Consecuencialismo es indirectamente contraproducente

La mayoría de mis afirmaciones, con pocos cambios, podrían incluir a todo un grupo de teorías morales. Estas son las diferentes versiones del *Consecuencialismo*, o C. La tesis central de C es

> (C1) Hay un fin moral último: que las consecuencias sean las mejores posibles.

C se aplica a todo. Aplicado a los actos, C establece tanto que

> (C2) Lo que cada uno de nosotros debe hacer es todo aquello que haga la consecuencia mejor, como que
> (C3) Si alguien hace lo que cree que producirá la consecuencia peor, está actuando mal.

Distinguí entre lo que tenemos más razón para hacer, y lo que sería para nosotros racional hacer, dado lo que creemos o debemos creer. Ahora tenemos que distinguir entre lo que es *objetiva* y *subjetivamente* correcto e incorrecto. Esta distinción no tiene nada que ver con la cuestión de si las teorías morales pueden ser objetivamente verdaderas. La distinción es entre lo que cierta teoría implica, dados (i) lo que son o habrían sido los efectos de lo que alguna persona hace o podría haber hecho, y (ii) lo que esta persona cree, o debe creer, acerca de estos efectos.

Puede ser útil mencionar una distinción similar. El tratamiento médico que es objetivamente correcto es aquel que sería el mejor para el paciente. El tratamiento que es subjetivamente correcto es aquel que, dada la evidencia médica, sería el más racional para que lo prescribiera el doctor. Como este ejemplo muestra, lo que sería mejor saber es lo que es objetivamente correcto. La parte central de una teoría moral contesta esta pregunta. Necesitamos una explicación de la corrección subjetiva por dos razones. A menudo no sabemos cuáles serían los efectos de nuestros actos. Y debemos ser censurados por hacer lo que está subjetivamente mal. Debemos ser censurados por tales actos aun si son objetivamente correctos. Un doctor sería censurado por hacer lo que muy probablemente matase a su paciente, aun si este acto, de hecho, salva la vida del paciente.

En lo que sigue casi siempre usaré *correcto*, *deber*, *bueno* y *malo* en el sentido objetivo. Pero *incorrecto* usualmente significará *subjetivamente* incorrecto, o *censurable*. El sentido que emplearé quedará claro a menudo por el contexto. Así, está claro que, de las tesis dadas arriba, (C2) trata acerca de lo que debemos objetivamente hacer, y (C3) trata sobre lo que es subjetivamente incorrecto.

Para cubrir casos de riesgo, C establece que

> (C4) Lo que subjetivamente debemos hacer es el acto cuyas consecuencias tienen la mayor bondad *esperada*.

Para calcular la bondad esperada de las consecuencias de un acto, el valor de cada posible efecto bueno se multiplica por la probabilidad de que el acto lo produzca. Lo mismo se hace con el disvalor de cada posible efecto malo. La bondad esperada de las conse-

cuencias es la suma de estos valores menos estos disvalores. Supongamos, por ejemplo, que si voy al Oeste tengo 1 entre 4 probabilidades de salvar 100 vidas, y 3 entre 4 probabilidades de salvar 20 vidas. La bondad esperada de mi ida al Oeste, valorada en términos del número de vidas salvadas es de $100 \times {}^{1}/_{4} + 20 \times {}^{3}/_{4}$, ó 25 + 15, ó 40. Supongamos a continuación que, si me voy al Este, salvaré 30 vidas con toda seguridad. La bondad esperada de mi ida al Este es de 30 x 1, ó 30. Según (C4), yo debo ir al Oeste, dado que el número esperado de vidas salvadas sería mayor.

El Consecuencialismo no sólo incluye actos y resultados, sino también deseos, disposiciones, creencias, emociones, el color de nuestros ojos, el clima y todo lo demás. Más exactamente, C incluye todo lo que pudiese hacer que las consecuencias fuesen mejores o peores. De nuevo usaré «motivos» para cubrir tanto los deseos como las disposiciones. C afirma que

> (C5) Los mejores motivos posibles son aquellos de los que es verdadero que, si los tuviéramos, las consecuencias serían las mejores.

Como antes, «posible» significa «causalmente posible». Y habría muchos conjuntos diferentes de motivos que serían los mejores en este sentido: no habría ningún otro conjunto posible de motivos del cual fuese verdadero que, si tuviéramos este conjunto, las consecuencias serían mejores. He descrito alguno de los modos en que podemos cambiar nuestros motivos. (C2) implica que debemos intentar hacernos tener, o mantener, cualquiera de los mejores conjuntos posibles de deseos. Para decirlo más en general, debemos cambiarnos a nosotros mismos, y también a cualquier otra cosa, de cualquier modo que haga mejores las consecuencias. Si creemos que está en nuestro poder producir tal cambio, (C3) implica que fracasar a la hora de hacerlo sería malo [9].

Con el fin de aplicar C, tenemos que preguntarnos qué es lo que hace a las consecuencias mejores o peores. La respuesta más simple nos la da el *Utilitarismo*. Esta teoría combina C con la siguiente tesis:

[9] Cf. Adams (2).

el mejor resultado es el que da a la gente la mayor suma neta de beneficios menos cargas, o, según la Versión Hedonista de la misma tesis, la mayor suma neta de felicidad menos dolor.

Hay muchas otras versiones de C. Estas pueden ser teorías *pluralistas*, que apelan a varios principios diferentes acerca de lo que hace a las consecuencias mejores o peores. Así, una versión de C apela tanto a la tesis utilitarista como al Principio de Igualdad. Este principio establece que es malo que a ciertas personas, sin que sea culpa suya, les vaya peor que a otras. Según esta versión de C, la bondad de una consecuencia depende tanto de cómo es de grande la suma neta de beneficios, como de si los beneficios y las cargas están distribuidos igualitariamente entre las diferentes personas. Un resultado podría ser mejor que otro, aunque conllevara una suma de beneficios menor, por la razón de que estos beneficios estuvieran repartidos más igualitariamente.

Un consecuencialista podría apelar a muchos otros principios. De acuerdo con tres de estos principios, es malo engañar, coaccionar y traicionar a las personas. Y algunos de estos principios pueden referirse esencialmente a sucesos pasados. Dos de tales principios apelan a derechos pasados, y a merecimientos justos. El Principio de Igualdad puede afirmar que las personas deberían recibir partes iguales, no en momentos determinados, sino en la totalidad que sus vidas constituyen. Si establece esta tesis, este principio se refiere esencialmente a sucesos pasados. Si nuestra teoría moral contiene tales principios, no sólo estamos preocupados con *consecuencias* en el sentido estrecho: con lo que ocurre *después de que* actuamos. Pero aún podemos ser, en un sentido más amplio, consecuencialistas. En este sentido más amplio nuestro fin moral último sería no que las consecuencias sean las mejores posibles, sino que la historia vaya lo mejor posible. Lo que digo abajo podría reformularse en estos términos.

Con la palabra «Consecuencialismo», y la letra «C», me referiré a todas estas diferentes teorías. Como pasaba con las diferentes teorías acerca del propio interés, haría falta como mínimo otro libro para decidir entre estas diferentes versiones de C. Este libro no discute esta decisión. Discuto sólo lo que estas diferentes versiones tienen en

común. Mis argumentos y conclusiones serían aplicables a todas, o a casi todas, las teorías plausibles de esta clase. Vale la pena enfatizar que, si un consecuencialista apela a todos los principios que he mencionado, su teoría moral es *muy* diferente del Utilitarismo. Como tales teorías rara vez han sido discutidas, esto es fácil de olvidar.

Algunos han pensado que, si el Consecuencialismo apela a principios muy diferentes, deja de ser una teoría distintiva, puesto que se la puede manejar para que incluya todas las teorías morales. Esto es un error. C recurre sólo a principios sobre lo que hace a las consecuencias mejores o peores. De modo que C podría afirmar que es peor si hubo más engaño o más coacción. C nos daría entonces a todos nosotros dos fines comunes. Debemos intentar hacer verdadero que haya menos engaño y menos coacción. Puesto que C les da a todos los agentes fines morales comunes, diré que C es *neutral con respecto al agente.*

Muchas teorías morales no toman esta forma. Estas teorías son *relativas al agente,* puesto que les dan a los diferentes agentes diferentes fines. Puede afirmarse, por ejemplo, que *cada quien* debería tener el fin de no coaccionar a los demás. Según esta concepción sería incorrecto por mi parte coaccionar a los demás, aun si al hacerlo pudiera dar lugar a que hubiera menos coacción. Afirmaciones similares podrían hacerse a propósito de engañar o traicionar a los otros. Según estas afirmaciones, el fin de cada persona debería ser, no que hubiera menos engaño o menos traición, sino que ella misma no engañara ni traicionara a los otros. Estas tesis no son consecuencialistas. Y son el tipo de tesis que la mayoría de nosotros acepta. C puede apelar a principios sobre el engaño y la traición, pero no apela a estos principios en su forma familiar.

Ahora describiré un modo diferente en que cierta teoría T podría ser contraproducente. Llamo a T

> *Indirecta y colectivamente contraproducente,* cuando ocurre que, si varias personas intentan lograr sus fines T-dados, estos fines serán logrados con más dificultad.

Según todas sus diferentes versiones, o por lo menos la mayoría, esto puede ser verdadero de C. C implica que deberíamos intentar hacer siempre todo aquello que tuviera las mejores consecuencias posibles. Si estamos dispuestos a actuar así, entonces somos *puros bienhechores.* Si todos fuéramos puros bienhechores, esto podría hacer que las consecuencias fuesen peores. Esto podría darse aun cuando siempre hiciéramos lo que, de los actos que fueran posibles para nosotros, produjese las mejores consecuencias. Los malos efectos podrían venir no de nuestros actos sino de nuestra disposición.

Hay muchos modos en que, si todos fuéramos puros bienhechores, esto podría tener malos efectos. Uno es el efecto sobre la suma de felicidad. Según cualquier versión plausible de C, la felicidad es una parte muy considerable de lo que hace mejores a las consecuencias. Y la mayor parte de nuestra felicidad viene de tener, y de obrar a partir de, ciertos deseos intensos. Estos incluyen los deseos que están implicados en el querer a otras personas, el deseo de trabajar bien, y muchos de los intensos deseos sobre la base de los que actuamos cuando no estamos trabajando. Para llegar a ser puros bienhechores, tendríamos que actuar contra la mayoría de estos deseos, o incluso suprimirlos. Y probablemente esto reduciría enormemente la suma de felicidad. Esto tendría peores consecuencias, aunque siempre hubiéramos hecho, de los actos que eran posibles para nosotros, lo que tenía mejores consecuencias. No podría hacer a las consecuencias peores de lo que *en realidad* son, dado cómo es efectivamente la gente. Pero haría a las consecuencias peores de lo que serían en caso de no ser nosotros puros bienhechores, en caso de que tuviésemos otros deseos y otras disposiciones causalmente posibles [10].

[10] Esto se argumenta en Sidgwick (1), pp. 431-9, y en Adams (2), de principio a fin. Adams discute lo que él llama *Utilitarismo de Actos y de Motivos,* afirmando que el Utilitarismo de Motivos no es sólo un caso especial del Utilitarismo de Actos. El utilitarista de motivos afirma que todo el mundo debería tener esas disposiciones que, al tenerlas, producirán el resultado mejor, aunque no haya nada que la persona pudiera haber hecho para determinarse a sí misma a tenerlas. Adams tiene razón cuando afirma que el Utilitarismo de Motivos es diferente del Utilitarismo de Actos, de la misma manera que el Utilitarismo de Reglas, al menos en ciertas versiones, no es un caso especial del Utilitarismo de Actos. No discutiré aquí estas distinciones.

Hay otras varias maneras en las cuales, si todos fuéramos puros bienhechores, esto podría producir el resultado peor. Una descansa en el hecho de que, cuando queremos actuar de cierto modo, probablemente nos engañaremos a nosotros mismos sobre los efectos de nuestros actos. Probablemente creeremos, de modo falso, que estos actos tendrán las mejores consecuencias. Considérese, por ejemplo, el matar a una persona. Si queremos que alguien esté muerto es fácil creer, falsamente, que esto produciría el resultado mejor. Por tanto lo que produce el resultado mejor es que estemos fuertemente dispuestos a no matar, aun cuando creamos que hacerlo produciría el resultado mejor. Nuestra disposición a no matar debería reemplazarse sólo cuando creamos que, al matar, produciríamos un resultado *muchísimo* mejor. Tesis similares se aplican al engaño, la coacción y otras varias clases de actos.

11. Por qué no falla C en sus propios términos

Asumiré que, así y de otros modos, C es indirecta y colectivamente contraproducente. Si todos fuéramos puros bienhechores, las consecuencias serían peores de lo que serían si tuviéramos otros conjuntos de motivos. Si sabemos esto, C nos dice que sería malo esforzarnos por ser, o por seguir siendo, puros bienhechores. Puesto que C hace esta afirmación, no está fallando en sus propios términos. C no se condena a sí misma.

Esta defensa de C es como mi defensa de PI. Pero es importante señalar una diferencia. PI es indirecta e individualmente contraproducente cuando es verdadero de alguna persona que, si nunca fuera abnegada, esto sería para ella peor de lo que sería si tuviera algún otro conjunto de deseos y disposiciones. Esto sería un mal efecto en los mismos términos de PI. Y este mal efecto ocurre a menudo. Hay muchas personas cuyas vidas van peor porque nunca son abnegadas, o lo son muy raramente. C es indirecta y colectivamente contraproducente cuando se cumple que, si algunos o todos nosotros fuéramos puros bienhechores, esto tendría peores consecuencias que las que tendría el que tuviéramos otros motivos distintos. Lo cual sería un mal efecto en términos de C. Pero este mal efecto puede *no* ocurrir. Hay pocos que sean puros bienhechores. Como son pocas tales personas, el hecho de que tengan esta disposición no puede producir peores consecuencias, tomadas las cosas en su conjunto.

El mal efecto en términos de PI ocurre con frecuencia. El mal efecto en términos de C puede no ocurrir. Pero esta diferencia no afecta a mi defensa de PI y C. Ambas teorías nos dicen que no tengamos las disposiciones que tendrían estos malos efectos. Por eso PI no falla, y C no fallaría, en sus propios términos. Es irrelevante si estos malos efectos ocurren en realidad.

Mi defensa de C asume que podemos cambiar nuestras disposiciones. Pero se puede objetar: «Supongamos que todos fuéramos puros bienhechores, porque creyéramos en C. Y supongamos que no pudiéramos cambiar nuestras disposiciones. Estas disposiciones tendrían malos efectos, en los términos de C, y estos malos efectos serían la consecuencia de nuestra creencia en C. De modo que C fallaría en sus propios términos». Hubo una objeción similar a mi defensa de PI. Discuto estas objeciones en la Sección 18.

12. La ética de la fantasía

He asumido que C es indirecta y colectivamente contraproducente. He asumido que, si todos fuéramos puros bienhechores, las consecuencias serían peores que las que se darían si tuviéramos otros conjuntos distintos de motivos. Si esta tesis es verdadera, C nos dice que debemos intentar tener uno de estos otros conjuntos de motivos.

El que esta tesis sea verdadera constituye en parte una cuestión fáctica. Creo que probablemente es verdadera. Pero no voy a intentar demostrarlo aquí. Parece que vale más la pena discutir lo que esta tesis implica. También creo que, aunque lleguemos a estar convencidos de que el Consecuencialismo es la mejor teoría moral, la mayoría de nosotros no se convertiría *de hecho* en un bienhechor puro.

Puesto que hace una asunción parecida, Mackie llama al Utilitarismo de Actos «la ética de la fantasía» [11]. Como otros autores, él asume que debemos rechazar una teoría moral si es en este sentido *exigente de modo no realista*: si ocurre que, aunque todos aceptáramos esta teoría, la mayoría de nosotros raramente haría lo que esta teoría afirma que debemos hacer. Mackie cree que una teoría moral es algo que nosotros *inventamos*. Si esto es cierto, sería plausible decir que una teoría aceptable no puede ser exigente de modo no realista. Pero, según otras varias concepciones de la naturaleza de la moralidad, esta tesis no es plausible. Podemos *esperar* que la mejor teoría no sea exigente de modo no realista. Pero, según estas opiniones, esto sólo puede ser una esperanza. No podemos asumir que tenga que ser verdadero.

Supongamos que estoy equivocado al asumir que C es indirecta y colectivamente contraproducente. Aunque sea falso, podemos asumir de modo plausible que C es exigente de modo no realista. Aunque tuviese peores consecuencias el que todos fuéramos bienhechores puros, probablemente sería causalmente imposible que todos o la mayoría de nosotros nos hiciéramos puros bienhechores.

Aunque estas sean asunciones muy diferentes, tienen la *misma* implicación. Si es causalmente imposible que nos convirtamos en puros bienhechores, C implica de nuevo que tenemos que intentar tener uno de los mejores conjuntos de motivos posibles, en términos consecuencialistas. Esta implicación es, por tanto, digna de discusión siempre y cuando (1) C sea o bien indirectamente contraproducente o bien exigente de manera no realista, o las dos cosas, y (2) ninguno de estos hechos demuestre que C no pueda ser la mejor teoría. Aunque todavía no estoy convencido de que C sea la mejor teoría, creo tanto en (1) como en (2).

13. CONSECUENCIALISMO COLECTIVO

Es importante distinguir C de otras formas de Consecuencialismo. Tal y como se ha formulado hasta aquí, C es *individualista* y se centra en efectos *reales*. Según C, *cada uno* de nosotros debería tratar de hacer lo que tuviera las mejores consecuencias, *dado lo que los otros van a hacer realmente*. Y cada uno de nosotros debería tratar de tener uno de los conjuntos posibles de motivos cuyos efectos fuesen mejores, dados los conjuntos reales de motivos que van a tener los otros. Cada uno de nosotros debe preguntar: «¿Hay algún otro conjunto de motivos que sea posible para *mí* y que, además, sea tal que, si *yo* tuviera este conjunto, las consecuencia serían mejores? Nuestras respuestas dependerían de lo que supiéramos, o pudiéramos predecir, acerca de los conjuntos de motivos que tendrán los otros.

¿Qué es lo que puedo predecir mientras mecanografío estas palabras, en el mes de enero de 1983? Sé que la mayoría de nosotros seguirá teniendo unos motivos muy parecidos a los que tenemos ahora. La mayoría de nosotros amará a ciertas personas, y tendrá los otros intensos deseos de los que la mayor parte de la felicidad depende. Como sé esto, C puede decir*me* que trate de ser un puro bienhechor. Esto puede dar el mejor resultado aun cuando, si *todos* fuéramos puros bienhechores, daría el peor resultado. Si las personas, en su mayoría, *no* fueran bienhechores puros, podría dar el mejor resultado si unas pocas personas lo fueran. Si la mayoría de la gente continúa siendo como es ahora, habrá mucho sufrimiento, mucha falta de igualdad, y muchas de las otras cosas que producen malos resultados. Buena parte de este sufrimiento yo podría fácilmente prevenirla, y con equidad, y también podría hacer mucho, de otras formas, para producir mejores consecuencias. Pueden por tanto favorecerse mejores consecuencias si evito los vínculos personales íntimos, y me las arreglo para debilitar comparativamente mis otros deseos intensos, de manera que pueda ser un puro bienhechor.

Con un poco suerte, tal vez no sea malo para mí transformarme de esta forma. Mi vida quedará despojada de la mayoría de las fuentes de felicidad. Pero una fuente de felicidad es la creencia de que uno actúa bien. Esta creencia me puede dar felicidad, haciendo de mi austera vida no sólo algo moralmente bueno sino también una buena vida para mí.

Puedo ser menos afortunado. Podría ocurrir que, aunque yo pudiera aproximarme a ser un puro bienhechor, esto no fuera una

[11] Mackie, capt. 4, sección 2.

buena vida para mí. Y podría haber muchas otras vidas posibles que fueran mucho mejores para mí. Cosa que podría ocurrir, según muchas de las teorías plausibles del propio interés. Las exigencias que me hace C pueden parecer por tanto injustas. ¿Por qué tendría que ser precisamente *yo* el que quitara a su vida la mayoría de las fuentes de felicidad? Más exactamente, ¿por qué debería estar yo entre los pocos que, siguiendo a C, debieran tratar de hacer esto? ¿No sería más equitativo que todos nosotros hiciéramos más para producir mejores consecuencias?

Esto sugiere una forma de Consecuencialismo que es *colectiva* y que además se centra en efectos *ideales*. Según esta teoría, cada uno de nosotros debería tratar de tener uno de los conjuntos de deseos y disposiciones que es tal que, si *todo el mundo* tuviera uno de estos conjuntos, esto daría mejores resultados que si todo el mundo tuviera otros conjuntos diferentes. Esta afirmación puede interpretarse de varios modos, y hay dificultades bien conocidas que se presentan a la hora de despejar las ambigüedades. Además, algunas versiones de esta teoría están abiertas a poderosas objeciones. Y es que nos dicen que ignoremos lo que de hecho ocurriría, en maneras que pueden ser desastrosas. Pero el Consecuencialismo Colectivo, o *CC*, tiene mucho atractivo. Más adelante sugeriré cómo una teoría más complicada podría mantener lo que resulta atrayente de CC, evitando al mismo tiempo las objeciones.

CC no se distingue de C sólo en sus tesis sobre nuestros deseos y disposiciones. Las dos teorías están en desacuerdo sobre lo que debemos hacer. Consideremos la cuestión de cuánto debería dar el rico al pobre. Para la mayoría de los consecuencialistas esta pregunta ignora los límites nacionales. Como yo sé que la mayoría de los otros ricos darán muy poco, será difícil para mí negar que lo mejor sería que yo regalara casi todo lo que ingreso. Incluso si diera nueve décimos, parte de mi décimo restante haría más bien si lo gastaran los muy pobres. De forma que el Consecuencialismo me dice que debo regalar casi todo lo que ingreso.

El Consecuencialismo Colectivo es mucho menos exigente. No me dice que dé la cantidad que en efecto produciría las mejores consecuencias. Me dice que dé la cantidad que es tal que si *todos* nosotros la diéramos, se producirían las mejores consecuencias. Más exactamente, me dice que dé lo que exigiese el particular impuesto internacional sobre la renta que daría el mejor resultado. Este impuesto sería progresivo, exigiendo aportaciones mayores de los que fuesen más ricos. Pero las exigencias que se le harían a cada persona serían mucho más pequeñas que las exigencias que hace C, según cualquier predicción plausible acerca de las cantidades que los otros deberán dar de hecho. Podría ser mejor que todos los que son tan ricos como yo dieran sólo la mitad de sus ingresos, o solamente un cuarto. Podría ocurrir que, si todos diésemos más, esto afectase tanto a nuestras economías que en el futuro tendríamos mucho menos para dar. Y podría ocurrir que, si todos diésemos más, nuestro donativo fuese demasiado grande para ser absorbido por las economías de los países más pobres.

La diferencia que he estado discutiendo surge sólo dentro de lo que se llama *teoría de la conformidad parcial*. Esta es la parte de una teoría moral que cubre casos en los que sabemos que algunas otras personas no harán lo que deben hacer. C podría requerir que unas pocas personas regalaran todo su dinero, e intentaran convertirse en bienhechores puros. Pero esto tendría sólo su causa en que la mayoría de las demás personas *no* hacen lo que C dice que deben hacer. No les dan a los pobres las cantidades que deberían dar.

En su teoría de la conformidad parcial, se ha dicho que C es demasiado exigente. No se trata de afirmar que C sea exigente *de manera no realista*. Como he dicho, creo que esto no supondría una objeción. Lo que se afirma es que, en su teoría de la conformidad parcial, C pone exigencias *injustas* o *no razonables*. Esta objeción se puede aplicar a *la teoría de la plena conformidad* de C. C sería mucho menos exigente si *todos* nosotros tuviéramos uno de los posibles conjuntos de motivos que, según C, debemos tratar de hacernos tener [12].

[12] El Consecuencialismo Colectivo no es el Consecuencialismo Cooperativo presentado en Regan, o la versión ampliada de C discutida en mi capt. 3.

Aunque C es indirectamente contraproducente, no falla en sus propios términos. Pero puede parecer expuesta a otras objeciones. Estas son como las que surgieron cuando discutimos PI. Supongamos que todos nosotros creemos en C, y que todos tenemos conjuntos de motivos que se hallan entre los mejores conjuntos posibles en términos consecuencialistas. He afirmado que, al menos para la mayoría de nosotros, estos conjuntos no incluirían ser un bienhechor puro. Si no somos puros bienhechores, a veces haremos lo que creemos producirá las peores consecuencias. Según C, entonces estaremos actuando mal.

Aquí tenemos un ejemplo. La mayoría de los mejores conjuntos de motivos posibles incluirían un gran amor hacia nuestros hijos. Supongamos que *Clara* tiene uno de estos conjuntos de motivos. Consideremos

> *El Caso Uno.* Clara podría o bien proporcionar a su hijo un beneficio, o bien dar beneficios mucho mayores a un desconocido infortunado. Puesto que ella ama a su hijo, le beneficia a él en lugar de beneficiar al desconocido.

Como consecuencialista, Clara puede asignar peso moral no sólo a cuánto se benefician los hijos, sino también a si son beneficiados *por sus propios padres.* Puede creer que el cuidado y el amor parentales son intrínsecamente, o en sí mismos, parte de lo que da mejores consecuencias. Aun así, Clara puede creer que, al no ayudar al desconocido, está produciendo una peor consecuencia. Por eso puede pensar que está actuando mal. Y este acto es absolutamente voluntario. Ella podría, si quisiera, evitar hacer lo que cree que es incorrecto. No lo hace así simplemente porque quiere beneficiar a su hijo más de lo que quiere evitar hacer el mal.

Si alguien hace libremente lo que cree que es incorrecto, queda expuesto usualmente a una crítica moral seria. ¿Debe Clara considerarse tan expuesta a semejante crítica? Como consecuencialista, podría negarlo. Su réplica sería como la de Kate cuando Kate afirmaba que ella no era irracional. Clara podría decir: «Actúo mal porque quiero a mi hijo. Pero sería malo para mí hacerme perder este amor. Que yo produzca el peor resultado es un efecto malo. Pero es parte de un conjunto de efectos que son, vistos en su totalidad, uno de los mejores conjuntos posibles. Sería por tanto malo para mí cambiar mis motivos de manera que yo no obrase mal en el futuro de esta manera concreta. Puesto que las cosas son así, cuando yo obro incorrectamente de esta manera, no es necesario que me considere *a mí misma* moralmente mala». Hemos visto que puede haber irracionalidad racional. Del mismo modo, puede haber *inmoralidad moral*, o maldad inocente. En tal caso, es el acto y no el agente el que es inmoral.

Una vez más, puede objetarse: «El mal efecto que Clara produjo podría haberse evitado. No es como el dolor que algunos cirujanos no pueden evitar causar cuando aplican el mejor tratamiento posible. El mal efecto fue el resultado de un acto separado y voluntario. Como podría haberse evitado, no puede decirse que sea parte de uno de los mejores conjuntos de efectos posibles».

Clara podría responder: «Yo podría haber actuado de modo diferente. Pero esto sólo significa que yo lo *habría* hecho así si mis motivos hubiesen sido diferentes. Dados mis motivos reales, es causalmente imposible que yo obrara de forma diferente. Y si mis motivos hubiesen sido diferentes, esto habría producido el peor resultado, consideradas las cosas en su conjunto. Como mis motivos reales son uno de los mejores conjuntos posibles, en términos consecuencialistas, los malos efectos *son*, en el sentido relevante, parte de uno de los mejores conjuntos de efectos posibles».

Puede objetarse: «Si no es causalmente posible que obres de forma diferente, dados tus motivos reales, no puedes hacer afirmaciones sobre lo que debes hacer. *Deber* implica *poder*».

Kate contestó a esta objeción en la Sección 6. No puede decirse que Clara debiera haber actuado de forma diferente si no hubiera podido haberlo hecho así. Esta última cláusula no significa «si esto hubiera sido causalmente imposible, dados sus motivos reales». Significa «si esto hubiera sido causalmente imposible, fuesen sus deseos los que fuesen».

Como Kate, Clara puede estar equivocada al asumir el determinismo psicológico. Si esto es así, sus afirmaciones pueden revisar-

se. Debería dejar de afirmar que, si tiene uno de los mejores conjuntos posibles de motivos, esto inevitablemente hará que ella haga lo que cree que es incorrecto. En vez de esto podría afirmar: «Si yo fuese una pura bienhechora, sería fácil no hacer lo que creo que es incorrecto. Como tengo otro conjunto de motivos, es muy difícil no actuar de este modo. Y sería incorrecto para mí cambiar mis motivos de modo que fuera más fácil no obrar de este modo. Como esto es así, cuando obro de este modo, soy moralmente mala nada más que en un sentido débil».

Consideremos a continuación

> *El Caso Dos.* Clara podría o bien salvar la vida de su hijo, o salvar las vidas de varios desconocidos. Puesto que ama a su hijo, le salva a él, y todos los desconocidos mueren.

En este caso, ¿podría Clara hacer las mismas afirmaciones? La muerte de varios desconocidos es un efecto muy malo. ¿Podría ella afirmar que estas muertes son parte de uno de los mejores conjuntos de efectos posibles? La respuesta puede ser No. Habría tenido una mejor consecuencia el que Clara no hubiera amado a su hijo. Esto habría sido peor para ella, y mucho peor para su hijo. Pero entonces habría salvado la vida de estos desconocidos. Este buen efecto podría haber tenido mayor peso que los malos efectos, produciendo un resultado mejor si consideramos las cosas en su conjunto.

Si esto es así, Clara podría decir: «No tenía razón alguna para creer que mi amor por mi hijo tendría este efecto tan malo. Fue subjetivamente correcto por mi parte permitirme querer a mi hijo. Y determinarme a perder este amor habría sido reprobable, o subjetivamente incorrecto. Cuando salvo a mi hijo en lugar de a los desconocidos, estoy actuando a partir de un conjunto de motivos que habría sido malo para mí determinarme a perder. Esto basta para justificar mi afirmación de que, cuando obro de este modo, esto es un caso de maldad inocente».

Podría decir un consecuencialista: «Cuando Clara se entera de que podría salvar a los desconocidos, *no* sería subjetivamente malo para ella determinarse a sí misma a no amar a su hijo. Esto sería bueno, puesto que entonces salvaría a los desconocidos». Pero Clara

podría responder: «Yo no podría haber perdido este amor con la celeridad que se hubiera requerido. Hay maneras en las que podemos cambiar nuestros motivos. Pero, en el caso de nuestros motivos más profundos, esto lleva mucho tiempo. *Habría* sido malo para mí tratar de perder mi amor por mi hijo. Si lo hubiera intentado, lo habría logrado sólo después de que los desconocidos hubieran muerto. Y esto hubiera tenido unas consecuencias todavía peores».

Como muestra esta réplica, las afirmaciones de Clara apelan esencialmente a ciertas asunciones fácticas. Podría haber sido verdadero que, si ella hubiera tenido la disposición de una pura bienhechora, esto habría tenido mejores consecuencias, vistas las cosas en su conjunto. Pero estamos asumiendo que esto es falso. Estamos asumiendo que el resultado será mejor si Clara tiene algún conjunto de motivos que a veces la harán elegir hacer lo que cree que tendrá las peores consecuencias. Y estamos asumiendo que su conjunto real de motivos es uno de los mejores conjuntos posibles.

Podríamos imaginar otros motivos que habrían tenido unas consecuencias aun mejores. Pero, dados los hechos acerca de la naturaleza humana, tales motivos no son causalmente posibles. Como Clara quiere a su hijo, ella le salva antes que a los desconocidos. Podríamos imaginar que nuestro amor por nuestros hijos se «desconectase» cada vez que las vidas de otras personas estuvieran en juego. Podría ocurrir que, si todos nosotros tuviéramos este tipo de amor, esto tendría mejores consecuencias. Si todos nosotros diéramos tal prioridad a salvar más vidas, habría pocos casos en los que nuestro amor por nuestros hijos tendría que desconectarse. Este amor podría por consiguiente ser tan grande como lo es ahora. Pero es de hecho imposible que nuestro amor pudiera ser como ese. No podríamos realizar semejante «ajuste fino». Cuando la vida de nuestros hijos estuviera amenazada, nuestro amor no podría desconectarse simplemente porque la de varios desconocidos se hallara también en peligro [13].

Clara alega que, cuando hace lo que cree que producirá el resultado peor, está obrando mal. Pero también afirma: «Porque obro

[13] Cf. Hare (2), p. 36.

sobre la base de un conjunto de motivos que sería malo para mí perder, estos actos son inocentes. Cuando obro de este modo, no es necesario que me considere mala. Si el determinismo psicológico no fuera verdadero, yo sería mala sólo en un sentido muy débil. Cuando actúo de este modo, no debería sentir remordimiento. Ni tampoco debería formar la intención de no obrar así otra vez».

Puede objetarse ahora que, puesto que hace estas afirmaciones, Clara no puede realmente *creer* que está obrando mal. Pero hay fundamentos suficientes para pensar que ella tiene esta creencia. Considérese el caso en que salva a su hijo en vez de a los desconocidos. Aunque ama a su hijo, Clara no cree que su muerte sería un resultado peor que la muerte de los desconocidos. Su muerte sería peor para él mismo y para ella. Pero la muerte de los desconocidos sería, vistas las cosas en su conjunto, mucho peor. Al salvar a su hijo y no a los desconocidos, Clara está haciendo lo que cree producirá un resultado mucho peor. Por tanto cree que está actuando mal. Su teoría moral implica directamente esta creencia. También cree que no debería sentir remordimiento. Pero la razón que tiene para creer esto no cuestiona su creencia de que está actuando mal. Su razón es que está actuando sobre la base de un motivo —el amor por su hijo— que habría sido malo para ella hacerse perder. Esto puede demostrar que no merece censura, pero no demuestra que no pueda creer que su acto es malo.

Puede decirse

(G4) Si alguien actúa sobre la base de un motivo que debe determinarse a tener, y que sería malo para él hacerse perder, no puede estar actuando mal .

Si (G4) estuviera justificado, apoyaría la afirmación de que el acto de Clara no habría sido malo. Y esto apoyaría la afirmación de que ella no puede realmente creer que su acto habría sido malo. Pero en la Sección 16 describo un caso en que (G4) no es plausible.

Clara podría añadir que, en muchos otros casos posibles, si ella creyera que su acto era malo, ella *creería* que ella misma era mala, y entonces sentiría remordimiento. Esto a menudo ocurriría si hiciera lo que creyera iba a tener las peores consecuencias, y *no* estuviera

actuando sobre la base de un conjunto de motivos que sería malo para ella hacerse perder. El Consecuencialismo por lo general no rompe la conexión entre la creencia de que un acto es incorrecto, y la censura y el remordimiento. La conexión se rompe sólo en casos especiales. Hemos estado discutiendo una de estas clases de caso: esos en que alguien obra sobre la base de un motivo que sería malo para él determinarse a perder.

Hay otra clase de caso en que se rompe la conexión. C se aplica a todo, incluyendo la censura y el remordimiento. Según C debemos censurar a otros, y sentir remordimiento, cuando esto tuviera las mejores consecuencias. Esto ocurriría cuando la censura o el remordimiento hiciesen cambiar nuestros motivos de un modo que significase la mejor consecuencia. Pero esto no ocurriría cuando, como Clara, tenemos uno de los mejores conjuntos de motivos posibles. Y podría no suceder aunque no tuviéramos tales motivos. Si somos censurados demasiado a menudo, la censura puede ser menos efectiva. C puede implicar así que, aunque no tengamos uno de los mejores conjuntos de motivos, deberíamos ser censurados sólo por aquellos actos de los que creamos que iban a tener un resultado *mucho* peor.

15. ¿PODRÍA SER IMPOSIBLE EVITAR ACTUAR MAL?

Las afirmaciones de Clara implican que ella no puede evitar hacer lo que cree que está mal. Podría decir: «No es causalmente posible *tanto* que tenga uno de los mejores conjuntos de motivos posibles, *como* que nunca haga lo que creo que está mal. Si fuera una pura bienhechora, mis actos corrientes nunca serían malos. Pero yo estaría actuando mal al permitirme a mí misma seguir siendo una pura bienhechora. Si en vez de eso me determino a mí misma a tener uno de los mejores conjuntos de motivos posibles, que es lo que debo hacer, entonces a veces haría lo que creo que está mal. Si no tengo la disposición de una pura bienhechora, no es causalmente posible que *siempre* obre como una pura bienhechora y nunca haga lo que creo que está mal. Como esto no es causalmente posible, y sería

malo para mí determinarme a mí misma a ser una pura bienhechora, no puedo ser moralmente criticada por no actuar siempre como una pura bienhechora».

Se puede decir ahora que, tal y como la describió Clara, C carece de uno de los rasgos esenciales de cualquier teoría moral. Puede objetarse: «Ninguna teoría puede exigir lo que es imposible. Como no podemos evitar hacer lo que C dice que está mal, no podemos hacer siempre lo que C dice que debemos hacer. Por eso deberíamos rechazar C. Como antes, *deber* implica *poder*».

Es de aplicación esta objeción aunque neguemos el determinismo psicológico. Supongamos que Clara hubiera salvado a su hijo en vez de a los desconocidos. Habría actuado de este modo porque ella no tiene la disposición de una bienhechora pura. Su amor por su hijo habría sido más fuerte que su deseo de evitar hacer lo que cree que está mal. Si negamos el determinismo, negaremos que, en este caso, habría sido causalmente imposible para Clara evitar hacer lo que cree que está mal. Por un esfuerzo de voluntad, ella podría haber actuado en contra de su deseo más fuerte. Aunque afirmáramos esto, no podríamos afirmar que Clara *siempre* pudiera actuar como una pura bienhechora *sin* tener una disposición de pura bienhechora. Aun los que niegan el determinismo no pueden romper completamente el enlace entre nuestros actos y nuestras disposiciones.

Si no podemos obrar siempre como puros bienhechores sin tener una disposición de un puro bienhechor, todavía es de aplicación la objeción mencionada arriba. Aunque neguemos el determinismo, tenemos que admitir lo siguiente. Asumimos que creemos verdaderamente que las consecuencias serían peores si todos nosotros fuéramos puros bienhechores. Si tenemos esta creencia, no es posible que nunca hagamos lo que creemos que va a tener un peor resultado. Si nos determinamos a nosotros mismos a ser, o nos permitimos seguir siendo, puros bienhechores, con ello hacemos lo que creemos tendrá un peor resultado. Si en vez de eso tenemos otros deseos y otras disposiciones, no será posible que siempre obremos como puros bienhechores, sin hacer jamás lo que creemos tendrá un peor resultado. El que pone la objeción puede por consiguiente decir: «Aunque el determinismo no sea verdadero, no es posible que

nunca hagamos lo que creemos tendrá peores consecuencias. Al afirmar que nunca debemos actuar de este modo, C exige lo que es imposible. Y puesto que deber implica poder, la tesis de C es indefendible».

Clara podría contestar: «En la mayor parte de los casos, cuando alguien obra mal merece ser censurado, y debería sentir remordimiento. Esto es lo que resulta más plausible en la doctrina de que deber implica poder. Es difícil de creer que pudiera haber casos en que, *sea lo que sea* lo que uno haga, o pudiera haber hecho con anterioridad, merezca ser censurado. C *no* implica esta creencia. Si salvara a mi hijo en vez de a los desconocidos, creería que estoy haciendo lo que tendrá unas consecuencias mucho peores. Creería por consiguiente que estoy obrando mal. Pero esto sería un caso de maldad *inocente*. Según C, nosotros *podemos* siempre evitar hacer *lo que merece ser censurado*. Esto basta para satisfacer la doctrina de que deber implica poder».

Podemos creer que estas afirmaciones no hacen frente a esta objeción con suficiente contundencia. Se planteó una objeción similar a PI. Es imposible que nunca hagamos lo que dice PI que es irracional. Empecé a contestar a esta objeción apelando al caso expuesto en la Sección 5: la Respuesta de Schelling al Robo a Mano Armada. En este caso, según cualquier teoría plausible de la racionalidad, yo no podría evitar actuar irracionalmente. Para hacer frente a la objeción a C, Clara podría apelar a otros casos en que no podemos evitar obrar mal. Que tales casos se dan, lo han afirmado algunos de los autores que se oponen con más decisión a C. Discuto esta respuesta en la siguiente nota [14].

[14] Primero deberíamos fijarnos en el sentido que tiene «poder» en la doctrina de que deber implica poder. Supongamos que se afirma que, en cierto caso, yo debía haber obrado de otra manera. Si yo no pudiera haber obrado de esta otra manera, no puede afirmarse que esto es lo que yo debía haber hecho. Como argumenté en la Sección 6, la afirmación (1) de que yo no podría haber obrado de esta otra manera no es la afirmación (2) de que obrar de esta manera habría sido imposible, dados mis reales deseos y disposiciones —o, para abreviar, mis motivos—. La afirmación es más bien (3) que obrar de esta manera habría sido imposible aunque mis motivos hubiesen sido diferentes. Obrar de esta manera habría sido imposible, *cualesquiera* que hubieran sido mis motivos.

16. ¿Podría ser correcto hacer que uno mismo obrara mal?

Como C es indirectamente contraproducente, nos dice que nos determinemos a nosotros mismos a hacer, o que nos las arregle-

A veces estamos en lo cierto al afirmar que deber implica poder. Supongamos que alguien cree que:

(A) Siempre está mal dejar de salvar la vida de una persona.

Supongamos que puedo salvar una vida o salvar otra, pero no puedo salvar las dos. Sea la que sea la vida que salve, dejo de salvar la vida de una persona. De acuerdo con (A), no puedo evitar obrar mal. Evitaría obrar mal sólo si salvara las dos vidas, y esto es imposible. Podemos rechazar de manera plausible (A), afirmando que deber implica poder. Yo no podría salvar estas dos vidas. No se trata de la afirmación de que esto no es posible dados mis motivos reales. Yo no podría salvar estas dos vidas, fueran los que fueran mis motivos. Como es en este sentido imposible para mí salvar las dos vidas, no puede afirmarse con justificación que esto es lo que debo hacer y que al dejar de salvar las dos vidas estoy obrando mal.

Volvamos ahora al Consecuencialismo. C afirma que es incorrecto para cualquiera hacer lo que él cree que producirá el resultado peor. Estamos asumiendo que, si todos fuéramos puros bienhechores, el resultado sería peor de lo que sería si tuviéramos otros motivos determinados. Si creemos esto, es imposible que nunca hagamos lo que creemos que producirá el resultado peor. Creeremos que, si somos puros bienhechores, hemos producido el resultado peor haciendo que nosotros mismos seamos, o permitiendo que sigamos siendo, puros bienhechores. Si no somos puros bienhechores pero tenemos los motivos que pensamos producirían el resultado mejor, no es posible que siempre actuemos como puros bienhechores —no haciendo nunca lo que pensamos que hará el resultado peor—. Con cualquiera de estas alternativas, es imposible que nunca obremos de esta manera .

¿Es imposible en el sentido que justifica una apelación a la doctrina de que deber implica poder? ¿Es imposible que nunca obremos de esta manera, sean los que sean o los que podrían haber sido nuestros motivos? Esto es verdadero, pero engañoso. Sugiere que esta imposibilidad no tiene nada que ver con cuáles sean nuestros motivos. Esto no es así. Esta imposibilidad implica esencialmente afirmaciones sobre nuestros motivos. ¿Por qué es imposible que nunca hagamos lo que creemos que producirá el resultado peor? Esto es imposible porque hay sólo una disposición, dada la cual sería causalmente posible no hacer *nunca* lo que creemos que va a producir el resultado peor, y hacer que nosotros mismos tengamos o mantengamos esta disposición sería *en sí mismo* un caso de hacer lo que creemos que va a hacer el resultado peor. Puesto que esta imposibilidad implica esencialmente

mos para que sea más probable hacer, lo que ella misma afirma que es moralmente incorrecto. Esto no es un defecto en los tér-

estas afirmaciones sobre nuestros motivos, no está claro que esta sea la clase de imposibilidad que justifica una apelación a la doctrina de que deber implica poder. Puede decirse al menos que este caso es muy diferente del caso en que es imposible para mí salvar las dos vidas. Esa imposibilidad no tenía nada que ver con mis motivos. En ese caso podíamos apelar convincentemente a la doctrina de que deber implica poder. Esto no demuestra que podamos apelar convincentemente a esta doctrina en este caso tan diferente. Tal vez podamos. Pero creo que no podemos apoyar aquí una apelación a esta doctrina afirmando que esta apelación es convincente en el caso en que no puedo salvar las dos de un par de vidas. Puesto que son tan diferentes, creo que ese caso es irrelevante.

Puede negarse la irrelevancia. Si creemos que el resultado sería peor si todos fuéramos puros bienhechores, es imposible que nunca hagamos lo que creemos que va a producir el resultado peor. He dicho que es engañoso afirmar que esto es imposible sean los que sean, o los que podrían haber sido, nuestros motivos. Esta afirmación sugiere, de manera falsa, que esta imposibilidad no tiene nada que ver con nuestros motivos. Pero, aunque sea engañosa, es todavía verdadera. Y esto podría decirse para hacer a este caso suficientemente similar a ese en el que no puedo salvar las dos vidas de un par de personas.

¿Hay casos en que podemos negar que deber implica poder, aunque la imposibilidad no tenga nada que ver con nuestros motivos? Algunos autores dicen que los hay. Nagel afirma que puede haber tragedias morales, casos en que, sea lo que sea lo que haga uno, siempre estará obrando mal. Nagel admite que, al hacer sus afirmaciones, está negando que *deber* implique *poder*. En los casos que describe, una persona debe evitar obrar mal, aunque no podría haber evitado obrar mal, cualesquiera que pudieran haber sido sus motivos. Es natural esperar que no pueda haber casos así. Pero Nagel escribe, «...es ingenuo suponer que hay una solución a todo problema moral con el que el mundo puede enfrentarnos». El mundo nos puede dar problemas para los cuales no hay solución que evite portarse mal (Nagel (2), p. 79).

Nagel sugiere que puede haber casos tales porque la mejor teoría moral contiene principios *contradictorios*. Podría afirmar: «Un principio moral singular no puede tener tales implicaciones. Si tal principio implica que no podemos evitar portarnos mal, este principio es insostenible». Esta afirmación es plausible cuando se aplica a casos como ese en que no puedo salvar las dos vidas de un par de personas. El Principio (A) implica que, al dejar de salvar las dos vidas, actúo mal. Aquí el fallo radica en este principio, no en el mundo.

Si es cierto que el resultado sería peor si todos fuéramos puros bienhechores, el Consecuencialismo implica que no podemos evitar obrar mal. Aunque el consecuencialista puede apelar a muchos principios morales diferentes, esta conclusión particular viene implicada por un principio singular. Viene implicada por la afir-

minos de C. Podemos hacer una pregunta como la que hice sobre la teoría del Propio Interés. C nos da un fin moral sustantivo:

mación de que es incorrecto hacer lo que creemos que va a producir el resultado peor. Pero, si puede haber casos en que no podemos evitar obrar mal, como sostiene Nagel, la explicación quizás no tenga que ser que hay un conflicto entre dos principios diferentes. La explicación no puede apelar simplemente al hecho de que es causalmente imposible actuar de dos maneras diferentes. Tiene que apelar a algo más profundo —algo como el conflicto entre dos principios diferentes—. Y, en el caso que estamos considerando, puede decirse que la explicación es de esta clase. Hay un conflicto, pero no entre dos principios diferentes, sino entre lo que sería el mejor conjunto de actos y lo que sería el mejor conjunto de motivos. Estamos asumiendo que es causalmente imposible, en vista de los hechos acerca de la naturaleza humana, *tanto* que realicemos el conjunto de actos que producirían el mejor resultado posible, *como* que tengamos esos motivos que, al tenerlos, producirían el mejor resultado posible. Este tipo de conflicto puede considerarse suficientemente similar al producido por el conflicto entre dos principios diferentes. Y puede decirse que aquí el fallo radica no en el principio de que es incorrecto hacer lo que creemos que va a producir el resultado peor, sino en el mundo. El consecuencialista podría repetir parte de la afirmación de Nagel. Podría decir: «puede ser cierto que, si tenemos la disposición que nos capacitará para no hacer nunca lo que creamos que va a producir el resultado peor, hemos producido el resultado peor al hacernos a nosotros mismos tener o mantener esta misma disposición. Si esto es así, el mundo nos ha dado un problema para el que no hay solución. Más exactamente, hay una solución, pero no nos permite evitar actuar mal. Hay algo que debemos hacer, bien mirado el asunto. Debemos determinarnos a nosotros mismos a tener, o a mantener, uno de los conjuntos de motivos que, al tenerlos, producirán un resultado mejor. Pero si tenemos uno de estos conjuntos, es imposible que no hagamos nunca lo que pensamos que va a producir un resultado peor. Si hacemos lo que, bien mirado, debemos hacer, es imposible que nunca actuemos incorrectamente».

Hay otra diferencia entre este caso y los que discute Nagel. Él sugiere que, en casos en que alguien no puede evitar obrar mal, se le debería culpar por obrar mal y debería sentirse culpable y arrepentirse. Esta es la afirmación que entra más agudamente en conflicto con la doctrina de que deber implica poder. Como he dicho, lo más difícil de creer es que, cualquier cosa que alguien haga o pudiera haber hecho, merezca ser culpado, y deba arrepentirse y sentirse culpable. El Consecuencialismo no implica esta afirmación. Cuando Clara salva a su hijo en vez de a los desconocidos, no merece ser culpada por ello, y no debe arrepentirse ni sentirse culpable. Quizás debiera sentir lo que Williams llama *pesar del agente*. Quizás debiera tener este sentimiento cuando piensa en los desconocidos muertos que podría haber salvado. Pero este sentimiento no es remordimiento ni culpabilidad.

que la historia vaya lo mejor posible. ¿Nos da también un segundo fin sustantivo: que nunca actuemos mal? Según la forma

C no implica que, cualquier cosa que Clara haga o pudiera haber hecho, merezca ser culpada y deba arrepentirse y sentirse culpable. Esto es lo que Williams y Nagel afirman de los agentes que se enfrentan a tragedias morales. Si pensamos que deber implica poder, podemos rechazar esta afirmación. Pero esto no sería una razón para rechazar C, desde el momento en que no implica esta afirmación.

Si un consecuencialista rechaza estos comentarios, tiene que revisar su teoría moral. Supongamos que admite que, si no pudiésemos bajo ningún concepto evitar siempre hacer lo que creemos que va a producir un resultado peor, no puede decirse que esto es lo que siempre debamos hacer. No puede decirse que sea siempre incorrecto hacer lo que creemos que va a producir un resultado peor. Como no es posible que no actuemos nunca de esta manera, no puede ser siempre incorrecto actuar de esta manera. Tiene que haber ciertos casos en que, aunque actuemos de esta manera, no estemos actuando incorrectamente. De acuerdo con (C3) es siempre incorrecto hacer lo que creemos que va a producir un resultado peor. Un consecuencialista podría abandonar ahora (C3) y sustituirlo por

> (C3') Siempre debemos evitar hacer lo que creemos que va a producir un resultado peor, *si esto es posible de un modo que en sí mismo no produzca un resultado peor*. Si esto es imposible, debemos evitar hacer lo que creemos que va a producir un resultado peor, siempre que pudiéramos haber obrado de forma diferente, de una manera que no habría producido un resultado peor.

Como he mantenido, «imposible» no significa aquí «causalmente imposible, dados nuestros motivos reales». Significaría «causalmente imposible, cualesquiera que pudiesen haber sido nuestros motivos». Si produciría un resultado peor el que todos fuésemos puros bienhechores, es causalmente imposible que no hagamos nunca lo que creemos que va a producir un resultado peor. Esto habría sido causalmente imposible, cualesquiera que pudiesen haber sido nuestros motivos. Como es causalmente imposible, en este sentido, que no actuemos *nunca* de esta manera, (C3') implica que, cuando actuamos de esta manera, no podemos estar actuando *siempre* incorrectamente. Si el Consecuencialismo afirma (C3') más bien que (C3), adopta una forma revisada que satisface la doctrina de que deber implica poder. Esta versión revisada de C anula la objeción que apela a esta doctrina.

¿Habría mucha diferencia si C se revisara de esta forma? Supongamos que todos aceptamos (C3'), y que de verdad creemos que el resultado sería peor si todos fuéramos puros bienhechores. El resultado sería mejor si tuviésemos uno de muchos otros conjuntos posibles de motivos. Supongamos además que es causalmente posible que nos determinemos a tener, o a permitirnos mantener, uno de estos otros conjuntos de motivos. (Sólo sobre la base de esta asunción surge esta

mejor conocida de C, el Utilitarismo, la respuesta es No. Para los utilitaristas, evitar la maldad es un simple medio para el logro del único fin moral sustantivo. No es en sí mismo un fin sustantivo. Y esto podría también afirmarse según las versiones de C que juzgan la bondad de las consecuencias en términos no de uno sino de varios principios morales. Podría afirmarse, por ejemplo, por parte de la teoría que apela tanto a la tesis utilitarista como al Principio de Igualdad. Todas estas teorías nos dan el fin *formal* de actuar moralmente, y evitar la maldad. Pero estas teorías podrían todas afirmar que este fin formal no es parte de nuestro fin moral sustantivo.

objeción a C.) (C3') implica que debemos determinarnos a nosotros mismos a tener, o a mantener, uno de estos otros conjuntos de motivos. Como es posible actuar en uno de estos sentidos, (C3') implica que sería incorrecto no hacerlo así.

Supongamos además que tenemos uno de estos otros conjuntos de motivos. Puesto que esto es así, a menudo hacemos lo que creemos que va a producir un resultado peor. (C3') implica que cuando actuamos de esta manera estamos actuando incorrectamente siempre que pudiéramos haber obrado de forma diferente, de una manera que no hubiera producido un resultado peor. ¿Habría sido posible que, en todos estos casos, obráramos de forma diferente, de un modo que no hubiera producido un resultado peor? No estamos preguntando aquí si esto habría sido causalmente posible, dados nuestros motivos reales. Estamos preguntando si esto habría sido posible, si nuestros motivos hubiesen sido diferentes. La respuesta a esta pregunta es No. No habría sido posible para nosotros haber actuado de forma diferente, en *todos* estos casos, de una manera que no hubiera producido un resultado peor. Podríamos haber actuado de forma diferente, en *todos* estos casos, sólo si todos fuéramos puros bienhechores, y esto habría producido un resultado peor.

Ahora deberíamos preguntar, «¿Podríamos haber actuado de forma diferente en *algunos* de estos casos, de una manera que no hubiera producido un resultado peor? La respuesta es Sí. Es imposible que obremos siempre como puros bienhechores sin tener la disposición de un puro bienhechor. Pero habría sido posible para nosotros haber obrado algunas veces de esta manera sin tener esta disposición. Si el Determinismo no está en lo cierto, podríamos haber actuado a menudo de esta manera. (C3') implicaría entonces que, aunque no siempre actuemos incorrectamente cuando actuamos de esta manera, a menudo actuamos incorrectamente. Al sustituir (C3) por (C3'), el consecuencialista anula la objeción que apela a la doctrina de que deber implica poder. Y esta revisión no supone mucha diferencia.

Aunque cualquier consecuencialista podría hacer esta afirmación, otras diversas teorías morales no la harían. Según estas teorías, la evitación de la maldad es en sí misma un fin moral sustantivo. Si aceptamos una de estas teorías, podemos plantear como mínimo dos objeciones a C. Podemos decir, «Una teoría aceptable no puede tratar el obrar moralmente como un mero medio». Esta objeción la discuto en la Sección 19. Podemos también decir, «Una teoría aceptable no puede decirnos que nos determinemos a nosotros mismos a hacer lo que esta misma teoría afirma que es incorrecto».

En caso de que planteáramos esta objeción, deberíamos preguntar si nosotros mismos creemos que los actos en cuestión serían incorrectos. Estamos considerando casos en que un consecuencialista cree que, aunque está obrando mal, él no es moralmente malo, porque obra según motivos que sería malo para él hacerse perder. En tales casos, ¿creeríamos que este consecuencialista está obrando mal?

Esto es improbable en el caso imaginario en que Clara salva a su hijo en vez de a los desconocidos. Si no somos consecuencialistas, probablemente creeremos que el acto de Clara no ha sido malo. Podemos pensar lo mismo sobre ciertos otros casos de esta clase. Supongamos que Clara se abstiene de matarme, aunque tiene la creencia verdadera de que matarme tendría el mejor resultado. Clara creería que, al abstenerse de matarme, estaría obrando mal. Pero consideraría esto como un caso de maldad inocente. Obra mal porque está fuertemente inclinada a no matar, y, por la razón dada al final de la Sección 10, ella cree que esta es una disposición que sería malo para ella hacerse a perder. Podemos creer una vez más que, al abstenerse de matarme, Clara *no* está obrando mal.

Si esta es nuestra opinión sobre estos casos, está menos claro que debiéramos plantear objeciones a esta parte de C. Aceptamos la afirmación de C de que, en estos casos, Clara no revelaría ser moralmente mala ella misma, ni merecedora de censura. Sobre esto no hay desacuerdo. Podemos objetar a la afirmación de C que, aunque Clara es inocente, sus actos son malos. Pero tal vez no deberíamos oponernos a esta afirmación, si no tiene sus implicaciones usuales.

Aún podemos objetar que una teoría moral aceptable no puede decirnos que nos las arreglemos para hacer lo que la misma teoría afirma que es malo. Pero consideremos

> *Mi Corrupción Moral.* Supongamos que tengo una carrera pública que se hundiría si se me implicara en un escándalo. Tengo un enemigo, un criminal a quien desenmascaré. Este enemigo, a quien ahora han puesto en libertad, desea vengarse. En vez de limitarse a dañarme, decide obligarme a corromperme, sabiendo que para mí esto será peor que la mayoría de los daños. Me amenaza con que él o algún miembro de su banda matará a todos mis hijos, a no ser que yo actúe de un modo obsceno mientras él lo filma. Si después mandara la película a un periodista, mi carrera quedaría hundida. Así que más adelante él podrá, amenazándome con destruir mi carrera, obligarme a obrar incorrectamente. Me obligará a ayudarle a cometer diversos delitos menores. Aunque yo soy moralmente tan bueno como la mayoría de la gente, la verdad es que no soy un santo. No obraría muy mal simplemente para salvar mi carrera; pero sí que ayudaría a mi enemigo a cometer delitos menores. En esto obraría mal, aún concediendo el hecho de que, si me negara a ayudar a mi enemigo, mi carrera quedaría arruinada. Podemos suponer además que, puesto que conozco bien a mi enemigo, tengo buenas razones para creer tanto que si yo me niego a dejarme filmar matará a mis hijos como que si no me niego, no los matará.

Debo dejar que este hombre haga su película. Podemos afirmar de manera plausible que *los gobiernos* no deberían ceder a semejantes amenazas, porque esto simplemente les expondría a amenazas ulteriores. Pero tal afirmación no sería válida para esta amenaza hecha por mi enemigo. Sería malo para mí rechazar su exigencia, con el resultado previsto de que mis hijos son asesinados. Debo dejarle hacer su película, aunque sepa que la consecuencia será que yo después obraré mal con frecuencia. Una vez asegurada la vida de mis hijos, a menudo ayudaré a mi enemigo, para salvar mi carrera, a cometer delitos menores. Estos actos posteriores serán totalmente voluntarios. No puedo afirmar que las posteriores amenazas de mi enemigo me fuercen a actuar así. Yo podría negarme a obrar mal, aunque ello suponga el hundimiento de mi carrera.

He afirmado que debo dejar que este hombre haga su película. En esto estarían de acuerdo incluso la mayoría de los que rechazan el Consecuencialismo. Estarían de acuerdo en que, dado que es la única manera de salvar la vida de mis hijos, yo debo hacer que ocurra que más tarde a menudo obre mal. De forma que estas personas creen que una teoría moral aceptable *puede* decirle a alguien que se determine a sí mismo a hacer lo que esta teoría dice que es incorrecto. Como ellos creen esto, no pueden objetar al Consecuencialismo que pueda tener esta implicación.

Si dejo a mi enemigo hacer su película, me estaría disponiendo a ayudarle a cometer delitos menores. Añadamos ahora algunos detalles a este caso. Yo podría obligarme a mí mismo a perder esta disposición, abandonando mi carrera. Pero mi enemigo me ha amenazado con que, si abandono mi carrera, su banda matará a mis hijos. Sería por tanto malo para mí obligarme a mí mismo a perder esta disposición. En contraste, si yo rechazo ayudar a mi enemigo a cometer sus delitos, simplemente destrozará mi carrera, mandando a un periodista la película en la que hago cosas obscenas. Mi enemigo me asegura que, si hunde mi carrera, mis hijos no serán asesinados. Obtiene un placer perverso al obligarme a hacer lo que sé que está mal, al amenazarme con hundir mi carrera. Perdería este placer si su amenaza fuera la de matar a mis hijos. Si le ayudase a cometer sus delitos porque este fuera el único modo de salvar la vida de mis hijos, yo no creería estar actuando mal. Pero como mi enemigo quiere que yo crea que estoy obrando mal, no me amenaza con *esto*.

Conociendo a mi enemigo, tengo buenas razones para creer en lo que dice. Como es la única manera de salvar la vida de mis hijos, debo dejarle hacer su película. Debo disponerme a ayudarle a cometer sus delitos menores. Y sería malo para mí determinarme a perder esta disposición, puesto que si lo hago asesinará a mis hijos. Pero, cuando actúo según esta disposición, estoy obrando mal. No debo ayudar a este hombre a cometer sus delitos simplemente para salvar mi carrera.

Este caso hace patente que deberíamos rechazar lo que llamé (G4). Esta es la afirmación de que, si debo obligarme a tener algu-

na disposición, y sería malo para mí obligarme a perder esta disposición, no puedo estar obrando mal cuando actúo a partir de esta disposición. En el caso recién descrito, cuando actúo a partir de una disposición semejante, *estoy* obrando mal [15].

Ahora expondré juntos cuatro errores similares. Algunos afirman que si es racional para mí obligarme a tener una cierta disposición, no puede ser irracional actuar sobre la base de esta disposición. Esto se reveló falso viendo el caso que llamé *la Respuesta de Schelling al Robo a Mano Armada*. Una segunda afirmación es que, si es racional para mí determinarme a creer que determinado acto es racional, este acto es racional. Esto se reveló falso viendo el caso que llamé *Mi Esclavitud*. Una tercera afirmación es que si hay determinada disposición que yo debo obligarme a tener, y que sería malo para mí obligarme a perder, no puede ser malo para mí obrar sobre la base de esta disposición. El ejemplo recién dado demuestra que esto es falso. Una cuarta afirmación es que, si debo determinarme a creer que cierto acto no sería incorrecto, este acto no puede ser incorrecto. Pronto demostraré que esto es falso. Estas cuatro afirmaciones asumen que la racionalidad y la corrección moral pueden *heredarse* o *transferirse*. Si es racional o correcto para mí o bien obligarme a estar dispuesto a actuar de cierto modo o bien hacerme a mí mismo creer que este acto es racional o correcto, este acto *es* racional o correcto. Mis ejemplos demuestran que esto no es así. La racionalidad y la corrección moral no pueden heredarse de este modo. En este respecto, la verdad es más simple que lo que estas afirmaciones implican.

17. CÓMO C PODRÍA SER MODESTA

Podría decirse que, si el Consecuencialismo rompiera a veces la conexión entre nuestra creencia de que nuestro acto es incorrecto

[15] (Nota añadida en 1987.) Mi ejemplo es defectuoso. No sería incorrecto por mi parte hacerme perder mi disposición, si lo hiciera mediante el rechazo a ayudar a mi enemigo, con el resultado de que él hunde mi carrera. El ejemplo podría repararse. (Podríamos suponer que él sólo dañaría mi carrera, y de este modo podría amenazar con daños ulteriores.)

y nuestra creencia de que nosotros somos malos, no continuaríamos en efecto considerando a la moralidad con la suficiente seriedad. De manera similar, nuestro deseo de evitar la maldad podría ser socavado si creyésemos que otros deseos serían con frecuencia más intensos. Este deseo puede sobrevivir sólo si creemos que debería ser *siempre* el primordial, y sentimos remordimiento cuando no lo es. Podría decirse, partiendo de esta base o de otras, que las consecuencias serían mejores si siempre mantuviéramos la conexión entre nuestras creencias morales y nuestras intenciones y emociones. Si esto fuera así, tendría mejores consecuencias que no creyésemos en C.

Pongo en duda estas afirmaciones. Pero vale la pena considerar lo que implicarían. Según C, cada uno de nosotros debería tratar de tener uno de los mejores conjuntos de deseos y disposiciones posibles, en términos consecuencialistas. Podría tener mejores consecuencias que no tuviéramos meramente estos deseos y disposiciones, sino que tuviéramos las emociones y las creencias morales correspondientes.

Consideremos, por ejemplo, el robo. Según algunas versiones de C, que la propiedad sea robada es intrínsecamente malo. Según otras versiones de C, esto no es así. Según estas versiones, el robo es malo sólo cuando produce las peores consecuencias. Evitar el robo no es parte de nuestro fin moral último. Pero podría ser verdadero que diese el mejor resultado el que estuviéramos fuertemente dispuestos a no robar. Y podría dar el mejor resultado el que creyéramos que robar es intrínsecamente malo, y sintiésemos remordimientos cuando robáramos. Podrían hacerse afirmaciones parecidas sobre muchos otros tipos de acto.

Si estas afirmaciones fuesen verdaderas, C sería modesta. Porque nos diría que debemos tratar de creer, no a ella misma, sino a alguna otra teoría. Debemos tratar de creer la teoría que es tal que, si la creyéramos, produciría las mejores consecuencias. Según las afirmaciones hechas arriba, esta teoría podría no ser C. Podría ser alguna versión de lo que Sidgwick llamó *Moralidad del Sentido Común*.

Si C nos dijera que creyéramos en alguna versión de esta moralidad, esto no sería Moralidad del Sentido Común como es ahora, sino una versión mejorada. La Moralidad del Sentido Común *no* es la teoría moral que, si creyésemos en ella, daría el mejor resultado. Tal teoría, por ejemplo, exigiría mucho más de los ricos. Podría dar el mejor resultado el que los que viven en las naciones más ricas dieran a los pobres al menos un cuarto o incluso la mitad de sus ingresos anuales. Los ricos ahora dan, y parece que creen que tienen toda la justificación para hacerlo, menos del uno por ciento.

Supongamos que C nos dijese que creyéramos en alguna otra teoría. Como he dicho, sería difícil cambiar nuestras creencias, si nuestra razón para hacerlo no es una razón que pone en duda nuestras viejas creencias, sino que es simplemente que tendría buenos efectos el tener diferentes creencias. Pero hay varios modos en que podríamos efectuar este cambio. Quizás podrían hipnotizarnos a todos, y la siguiente generación podría ser educada de diferente forma. Tendrían que hacernos olvidar cómo y por qué adquirimos nuestras nuevas creencias, y el proceso habría que ocultárselo a los futuros historiadores.

Supondría una diferencia el que aceptáramos, no C, sino el Consecuencialismo Colectivo. Si aceptamos C podríamos concluir que C la debe rechazar la mayoría de la gente, pero aún unas pocas personas deberían creer en ella. Nuestra teoría sería entonces en parte modesta, y en parte *esotérica*, y debería decir a los que en ella creen que no ilustraran a la mayoría ignorante. Pero como consecuencialistas colectivos deberíamos creer en la teoría moral que es tal que, si *todos* creyéramos en ella, se darían las mejores consecuencias. Esta teoría no puede ser esotérica.

Algunos encuentran especialmente inaceptable que una teoría moral pueda ser esotérica. Si creemos que el engaño es moralmente incorrecto, el engaño acerca de la moralidad puede parecer especialmente incorrecto. Sidgwick escribió: «parece conveniente que la doctrina de que la moral esotérica es conveniente debería mantenerse esotérica. O bien, si esta ocultación resulta difícil de mantener, puede ser deseable que el Sentido Común deba repudiar las doctrinas que es conveniente confinar a una minoría ilustrada» [16]. Esto es lo que Williams denomina Consecuencialismo del «Palacio del Gobernador», puesto que trata a la mayoría como a los nativos en una colonia [17]. Como dice Williams, no podemos dar la bienvenida a una conclusión semejante. Sidgwick lamentó sus conclusiones, pero no pensó que lamentarse fuese una razón para la duda [18].

He afirmado que es improbable que C sea totalmente modesta. Como mucho, sería en parte modesta y en parte esotérica. Podría tener las mejores consecuencias el que algunas personas no creyeran en C; pero es improbable que produjese el mejor resultado el que nadie en absoluto creyese en C.

Aquí tenemos otra razón para poner esto en duda. Supongamos que todos nosotros llegamos a creer en C. (Esto parecerá menos inverosímil si recordamos que C puede ser una teoría pluralista, que apele a muchos principios morales diferentes.) Entonces decidimos que C es completamente modesta. Decidimos que produciría las mejores consecuencias el que nos resolvamos a creer alguna versión mejorada de la Moralidad del Sentido Común. Podríamos tener éxito en producir este cambio en nuestras creencias. Dados ciertos cambios en el mundo, y en nuestra tecnología, podría suceder después que tuviese las mejores consecuencias el que revisáramos nuestras creencias morales. Pero si ya no creyéramos más en C, porque ahora creyéramos en alguna versión de la Moralidad del Sentido Común, no seríamos llevados a hacer estas necesarias revisiones en nuestra moralidad. Nuestra razón para creer en ella no sería que *ahora* creamos que ella es la moralidad que tendría las mejores consecuencias si creyésemos en ella. Esta sería la razón por la cual nos resolvemos a creer en esta moralidad. Pero, para creer en ella, tenemos que haber olvidado que esto es lo que hicimos. Ahora simplemente creeríamos en ella. Podríamos por tanto no ser llevados a revisar nuestra moralidad aunque llegase a suceder que nuestra cre-

[16] Sidgwick (1), p. 490.

[17] Sen y Williams, p. 16.

[18] «Siendo consciente de que la verdad más profunda que tengo que revelar no tiene nada que ver con las «buenas noticias»...yo no diría nada, si pudiera... que hiciera a la filosofía —mi filosofía— popular». Sidgwick (2), pp. 395-6.

encia en ella incrementara las probabilidades de algún desastre, como por ejemplo una guerra nuclear.

Estas afirmaciones deberían afectar a nuestra respuesta a la pregunta de si daría el mejor resultado el que dejáramos de creer en C. Podríamos creer correctamente que hay alguna otra teoría moral que tuviera las mejores consecuencias si creyésemos en ella. Pero una vez que el Consecuencialismo se ha puesto a sí mismo en segundo plano, y se han soltado amarras, las consecuencias a largo plazo podrían ser mucho peores.

Esto sugiere que, como mucho, lo que podría suceder es que C fuera en parte modesta. Podría ser mejor si la mayoría de la gente se resolviera a creer alguna otra teoría por algún proceso de autoengaño que, para tener éxito, también tendría que ser olvidado. Pero, como precaución, unas cuantas personas deberían seguir creyendo en C, conservando además evidencia convincente sobre este autoengaño. Estas personas no haría falta que vivieran en el Palacio del Gobernador, ni que tuvieran ningún otro estatus especial. Si las cosas fueran bien, esos pocos no tendrían que hacer nada. Pero si la teoría moral en que la mayoría creyera se hiciera desastrosa, los pocos podrían entonces revelar lo que saben. Cuando la mayoría de la gente cayera en la cuenta de que sus creencias morales eran el resultado del autoengaño, esto socavaría estas creencias, y prevendría el desastre.

Aunque he dicho que esto es improbable, supongamos que C *fuera* completamente modesta. Supongamos que esta teoría nos dijera que todos nos resolviéramos a creer no en ella misma sino en alguna otra teoría. Williams dice que, si esto fuera así, la teoría dejaría de merecer su nombre, puesto que «no determina para nada cómo se comporta el pensamiento en el mundo» [19]. Esta afirmación es desconcertante puesto que, como también dice Williams, C exigiría que el modo en que pensamos acerca de la moralidad, y nuestro conjunto de deseos y disposiciones, «tiene que ser para lo mejor» [20]. Esto es exigir algo bastante específico, y completamente consecuencialista.

[19] Williams, p. 135; Sen y Williams, p. 15.
[20] Williams, p. 135; Sen y Williams, p. 15.

Williams hace la tercera afirmación de que, si C fuera completamente modesta, dejaría de ser efectiva [21]. Esto no es necesariamente así. Supongamos que las cosas suceden como se describió arriba. Todos pasamos a creer en alguna variante de C. Entonces pasamos a creer verdaderamente que, si todos creyéramos en alguna otra teoría, esto produciría el mejor resultado posible. C nos dice a todos que creamos en esta otra teoría. De algún modo indirecto, efectuamos este cambio en nuestras creencias. Nadie cree ahora en C. Pero esto no justifica la afirmación de que C ha dejado de ser efectiva. Ha tenido el efecto de que todos nosotros creemos ahora en alguna otra teoría particular. Y nuestra creencia en esta otra teoría producirá las mejores consecuencias. Aunque nadie cree en C, C es todavía efectiva. Hay dos hechos persistentes que son los efectos de nuestra anterior creencia en C: nuestras nuevas creencias morales, y el hecho de que, puesto que tenemos estas creencias, el resultado es todo lo bueno que puede ser.

Williams establece correctamente que, si C fuera modesta del todo, no estaría claro lo que esto demuestra. Tendríamos que decidir si se demostraba que C «es inaceptable, o meramente que nadie debe aceptarla» [22]. Está claro que, a la vista de nuestras últimas asunciones, nadie debe moralmente aceptar C. Si alguien aceptara C, ella misma le diría que debe moralmente tratar de rechazar C y, en vez de en ella, creer en alguna otra teoría. Pero, como sugiere Williams, hay dos preguntas. Una pregunta es la de si hay alguna teoría que sea la única en que *debemos moralmente* intentar creer. Otra pregunta es la de si esta es la teoría en que *debemos* creer *intelectualmente* o *en términos de búsqueda de la verdad* —si esta teoría es la teoría verdadera o la mejor justificada—. Afirmé antes que, si una teoría de la racionalidad fuese modesta, esto no demostraría que esta teoría no pudiese ser la teoría verdadera o la mejor justificada. ¿Podemos sentar una afirmación similar para las teorías morales?

Nuestra contestación a esta pregunta dependerá en parte de nuestras creencias acerca de la naturaleza del razonamiento moral.

[21] Williams, p. 135; Sen y Williams, p. 15.
[22] Williams, p. 135; Sen y Williams, p. 15.



Si una teoría moral puede ser *verdadera* en un sentido totalmente franco, está claro que, si es modesta, esto no demuestra que no pueda ser verdadera. Pero podemos, en vez de esto, considerar a la moralidad como un producto social, o bien en el sentido real o de algún modo «constructivista ideal». Entonces podemos decir que, para ser aceptable, una teoría moral tiene que cumplir lo que Rawls llama «la condición de la publicidad»: tiene que ser una teoría que todos deban aceptar, y reconocer públicamente de forma mutua [23]. Según estas concepciones metaéticas, una teoría moral no puede ser modesta. Según concepciones distintas, lo puede ser. Llevaría como mínimo un libro decidir entre estas concepciones diferentes. Por eso tengo que dejar, en este libro, esta cuestión abierta. Esto no importaría si, como creo, C *no* fuese modesta.

18. La objeción que asume inflexibilidad

Volveré ahora a una objeción que se presentó anteriormente. Consideremos a esas personas para las que la teoría del Propio Interés es indirectamente contraproducente. Supongamos que estas personas creen en PI, y en consecuencia no son nunca abnegadas. Esto es peor para ellas. Sería mejor para ellas si tuvieran otros deseos y otras disposiciones. En el caso de estas personas, esto no sería posible como no creyeran en una teoría diferente. Y podría suceder que no pudieran cambiar ni sus creencias ni sus disposiciones.

Afirmaciones similares podrían ser verdaderas para los consecuencialistas. Supongamos que, porque todos nosotros creemos en C, somos todos puros bienhechores. Esto produce un resultado peor del que produciría el que todos nosotros tuviéramos otras disposiciones. Pero no podemos cambiar nuestras disposiciones a no ser que también cambiemos nuestras creencias acerca de la moralidad. Y no podemos llevar a cabo estos cambios.

Es improbable que todas estas afirmaciones sean verdaderas. Si lo fuesen, ¿plantearían objeciones a PI y a C?

[23] Rawls, pp. 133 y 182, sobre todo la nota a pie de página 31.

Puede servir de ayuda el considerar un caso imaginario. Supongamos que Satán gobierna el universo. Satán no puede influir en cuál es la teoría verdadera de la racionalidad, ni en cuál es la teoría mejor o mejor justificada. Pero él sabe cuál es esta teoría, y perversamente hace que la creencia en ella tenga malos efectos en sus propios términos. Al imaginar este caso no tenemos necesidad de asumir que la mejor teoría es la teoría del Propio Interés. Sea cual sea la mejor teoría, Satán haría que la creencia en ella tuviera malos efectos, en los términos de esta teoría.

A continuación podemos asumir lo mismo sobre las teorías morales. Supongamos que la mejor teoría moral es el Utilitarismo. Según esta teoría, todos deberíamos tratar de producir el resultado que sea el mejor para todos, imparcialmente considerado. Satán se asegura de que, si la gente cree en esta teoría, esto es peor para todos. Supongamos a continuación que la mejor teoría moral no es consecuencialista, y que manda a cada persona no engañar nunca a otras, ni coaccionarlas, ni tratarlas injustamente. Satán se asegura de que los que crean en esta teoría serán de hecho, a pesar de sus intenciones en sentido contrario, más engañadores, coartadores e injustos.

Satán se asegura de que, si alguien cree en alguna teoría, esto tenga malos efectos en sus propios términos. ¿Esto haría algo para demostrar que tal teoría no es la mejor teoría? Está claro que no. Como mucho, lo que podría demostrarse es que, dada la interferencia de Satán, sería mejor si nosotros no creyéramos en la mejor teoría. Puesto que somos meros juguetes de Satán, la verdad acerca de la realidad es extremadamente deprimente. Podría ser mejor que tampoco conociéramos esta verdad.

En este caso imaginario, sería mejor que no creyéramos en la mejor teoría. Esto muestra que deberíamos rechazar

> (G5) Si debemos determinarnos a creer que determinado acto es incorrecto, este acto *es* incorrecto.

Como afirmé, la incorrección no puede *heredarse* de esta manera. Supongamos que todo lo que sabemos es que la creencia en

determinada teoría tendría malos efectos, en los propios términos de esta teoría. Esto no demostraría que esta no es la mejor teoría. Si esto es cierto tiene que depender de *por qué* la creencia en la teoría tiene estos malos efectos. Hay dos posibilidades. Los malos efectos pueden ser producidos porque hemos hecho exitosamente lo que esta teoría nos dice que hagamos. Si esto fuera verdad, la teoría sería *directamente* contraproducente, y esto podría refutarla. Los malos efectos pueden, en cambio, ser producidos por algún hecho muy diferente acerca de la realidad. Si este hecho fuera la interferencia de Satán, esto no cuestionaría la teoría.

¿Qué deberíamos decir sobre las posibilidades descritas arriba? Supongamos que lo que sigue es verdadero. Sería peor para cada uno de nosotros si creyéramos en PI, y por tanto nunca fuéramos abnegados. Si todos nosotros creyéramos en C, y por eso fuéramos puros bienhechores, esto daría un resultado peor. Y, si tuviéramos alguna de estas creencias y disposiciones, seríamos incapaces de cambiarlas. Entonces sería verdadero que la creencia en estas dos teorías tendría malos efectos en los términos de las propias teorías. ¿Esto las cuestionaría? ¿O sería simplemente como la interferencia de Satán?

Puede que la mejor teoría no sea ni PI ni C. Yo defenderé más adelante que debemos rechazar PI. Pero si me equivoco, y o PI o C es la mejor teoría, sugiero que las posibilidades recién descritas no supondrían objeción alguna para ninguna de las teorías. Si o PI o C es la mejor teoría, la creencia en esta teoría tendría malos efectos, en los propios términos de esta teoría. Pero estos malos efectos no serían el resultado de hacer por nuestra parte, o tratar de hacer, lo que PI o C nos dicen que hagamos. Los malos efectos serían el resultado de nuestras disposiciones. Y estas teorías no nos mandarían tener estas disposiciones. Nos mandarían que, si podemos, no las tengamos. PI nos mandaría que, si podemos, no nos neguemos a ser abnegados. C nos mandaría que, si podemos, no seamos puros bienhechores. Tendríamos una de estas disposiciones porque creemos en una de estas teorías. Pero estas teorías no nos dicen que creamos en ellas mismas. PI le dice a cada persona que crea en la teoría que sería mejor para ella creer. C nos dice que creamos en

la teoría que produciría las mejores consecuencias si creyéramos en ella. Según las asunciones hechas arriba, PI y C *no* nos mandarían creer en PI y en C.

Puesto que creemos en PI o en C, la creencia en cualquiera de las dos tendría malos efectos en los términos de esta teoría. Pero estos malos efectos *no* serían la consecuencia de que hiciéramos lo que estas teorías nos dicen que hagamos. Serían la consecuencia de que tenemos disposiciones que estas teorías *no* nos dicen que tengamos, y que nos dicen que, si podemos, *no* tengamos. Y serían la consecuencia de que creemos en lo que estas teorías *no* nos dicen que creamos, y que nos dicen que, si podemos, *no* creamos. Puesto que no podemos culpar a estas teorías de estos malos efectos en ninguno de estos sentidos, sugiero que, si las afirmaciones hechas arriba fueran verdaderas, esto no pondría en duda estas teorías. Estas afirmaciones serían meramente, como la interferencia de Satán, verdades deprimentes sobre la realidad.

19. ¿Puede el ser racional o moral ser un simple medio?

PI nos dice que actuemos racionalmente, y C nos dice que actuemos moralmente. Pero estos son sólo lo que yo llamo nuestros fines formales. He asumido que actuar moralmente no sería como tal un fin sustantivo que C nos dé. C podría afirmar que actuar moralmente es un mero medio. De forma similar, actuar racionalmente puede no ser parte del fin sustantivo que nos es dado por PI. Y PI podría afirmar que actuar racionalmente es un mero medio. ¿Constituye esto una objeción a estas dos teorías?

Hay una diferencia aquí entre PI y C. PI no puede afirmar que nuestro fin formal sea como tal un fin sustantivo. Pero C *podría* hacer esta afirmación. Puede haber aquí una objeción a PI. Pero no puede haber una objeción similar a C.

Podríamos objetar, a PI: «Si actuar racionalmente no es un fin que debiéramos tener, sino un mero medio, ¿por qué deberíamos ser racionales? ¿Por qué deberíamos querer saber lo que tenemos más razón para hacer?

¿De qué modo debería contestar a esto un teórico del Propio Interés? Podría aceptar la Teoría de la Lista Objetiva, en relación con el propio interés. Podría de este modo afirmar: «Ser racional y actuar racionalmente son, en sí mismos, parte de lo que hace que nuestras vidas marchen mejor. Si son, vistas las cosas en su conjunto, mejores para nosotros, PI no implica que ser racional y actuar racionalmente sean meros medios. Son, en sí mismos, partes del fin último de cada persona PI-dado».

Consideremos a continuación a un teórico del Propio Interés que sea hedonista. Esta persona tiene que admitir que cree que actuar racionalmente es un mero medio. Pero podría decir: «Según PI, lo que tienes más razón para hacer es todo aquello que haga que tu vida marche lo mejor posible. Si quieres saber lo que tienes más razón para hacer, y quieres actuar racionalmente, PI no implica que estos sean deseos sin importancia. Esto no está implicado por la afirmación de que, si tú sigues PI, y actúas racionalmente, tus actos importan meramente como medio. Importar como medio es una manera de importar. Tus deseos no serían importantes sólo si actuar racionalmente no importara. PI afirma que, cuando estás decidiendo qué hacer, comparado con actuar racionalmente, *nada importa más*. Esta última afirmación está justificada aun cuando lo que sería racional que hicieras fuese disponerte a actuar irracionalmente. Lo que importa *más*, aun en este caso, es que hagas lo que sería racional que hicieras».

Me vuelvo ahora de PI a C. C podría afirmar que actuar moralmente es un mero medio. Podemos objetar: «Si es así, ¿por qué deberíamos preocuparnos por la moralidad? Consideremos primero la forma más simple de consecuencialista, el utilitarista hedonista. Semejante persona podría decir: «Importa moralmente si lo que ocurre es bueno o malo. Es malo que haya más sufrimiento, bueno que haya más felicidad. También importa que actuemos moralmente y evitemos la maldad. Deberíamos tratar de hacer lo mejor que podamos para reducir el sufrimiento e incrementar la felicidad. Esto importa, no en sí mismo, sino a causa de sus efectos. En este sentido, evitar la maldad es un mero medio. Pero esto no implica que *no* importe moralmente que evitemos la maldad. Cuando estamos

decidiendo qué hacer, comparado con evitar la maldad, *nada importa más*. Esta última afirmación está justificada aun cuando lo que debamos hacer sea adquirir la disposición de actuar incorrectamente. Lo que importa *más*, aun aquí, es que hagamos lo que debemos hacer».

Un utilitarista hedonista tiene que admitir que, según su concepción, *si* la maldad no tuviera malos efectos, no importaría. Si hubiera más maldad, esto no produciría en sí mismo consecuencias peores. Como con la versión hedonista de PI, el logro de nuestro fin formal importa, pero sólo como medio.

¿Puede defender su afirmación este utilitarista? Podría primero apelar a la falta de atractivo de lo que Williams denomina *autoindulgencia moral* [24]. Comparemos a dos personas que intentan aliviar el sufrimiento de los demás. La primera persona actúa porque simpatiza con ellos. También cree que el sufrimiento es malo y debe ser aliviado. La segunda persona actúa porque quiere pensar de sí misma como en alguien que es moralmente bueno. De estas dos personas, la primera parece ser mejor. Pero la primera persona no tiene pensamientos acerca de la bondad de actuar moralmente, o de la condición mala de la maldad. Es movida a actuar simplemente por su simpatía, y por su creencia de que, como el sufrimiento es malo, debe intentar prevenirlo. Esta persona parece considerar que actuar moralmente es un mero medio. Es la segunda persona la que considera actuar moralmente como un fin por derecho propio que es en sí mismo bueno. Como la primera persona parece ser mejor, esto apoya la afirmación de que actuar moralmente *es* un mero medio.

Consideremos a continuación

> *El Asesinato y la Muerte Accidental*. Supongamos que yo sé que X está a punto de morir, y que, en lo que será su última actuación, intenta asesinar a Y. Y también sé que, a no ser que se rescate a Z, morirá a causa de un incendio forestal. Yo podría ser capaz de persuadir a X de que no asesinara a Y, o ser capaz en vez de eso de salvar la vida de Z. Supongamos que creo que, si Y fuese asesinado por X, esto no sería peor que el que Z muriera en un incendio forestal. Estos dos resultados serían igualmente malos, porque cada uno de ellos

[24] Williams (6), capt. 3.

involucraría que una persona muere. En el primer resultado habría además un caso muy serio de maldad. Pero, según mi teoría, esto no produciría un resultado peor. (Podría ser al contrario si el malvado sobreviviese. Pero X está a punto de morir.) Puesto que creo que la maldad, como tal, no produce un resultado peor, creo que, si mi probabilidad de salvar a Z fuera *ligeramente* más alta que mi probabilidad de convencer a X de que no asesinara a Y, yo debería intentar salvar a Z [25].

Muchos aceptarían esta última conclusión. Creerían que, si mi probabilidad de salvar a Z es ligeramente más elevada, yo debería intentar salvarle a él antes que a Y. Si aceptamos esta conclusión, ¿podemos afirmar también que es malo en sí mismo que haya más maldad? ¿Podemos afirmar que debo intentar prevenir la muerte accidental de Z aunque, como implica maldad, el asesinato de Y a manos de X produciría una consecuencia peor? Sería difícil defender esta afirmación. Para muchos, este es otro caso que apoya la concepción de que la evitación de la maldad es un mero medio.

Pero puede objetarse: «Si X intenta asesinar a Y, la maldad está ya presente. No evitas la maldad convenciendo simplemente a X de que no lleve a cabo su intención asesina. Esta es la razón por la cual tú debes tratar de salvar a Z». Podemos cambiar el ejemplo. Supongamos que yo sé que X puede pronto dar en pensar falsamente que ha sido traicionado por Y. X no es moralmente malo, pero es como Otelo. Es bueno, pero potencialmente malo. Sé que es probable que, si X cree que ha sido traicionado por Y, asesinará a Y, como Otelo. Yo tengo una buena probabilidad de evitar que X llegue a adquirir esta falsa creencia, y evitar así que asesine a Y. Pero tengo una probabilidad ligeramente más elevada de salvar la vida de Z. Como antes, X se halla de cualquier modo a punto de morir. Muchos creerían una vez más que debo tratar de salvar la vida de Z. Esto sugiere que, si X forma su intención, y entonces asesina a Y, no se produciría una consecuencia peor que si Z muriera accidentalmente. Si esto *produjera* una consecuencia peor, ¿por qué debería yo tratar de salvar a Z antes que a Y? ¿Por qué debería tratar de evi-

[25] Debo este ejemplo a T. Scanlon.

tar el menor de dos males, cuando mi probabilidad de éxito es sólo ligeramente mayor?

Supongamos que, puesto que creemos que yo debo tratar de salvar a Z, estamos de acuerdo en que la muerte de Y no tendría una consecuencia peor que la de Z. Si esto es así, lo que hay de malo en la muerte de Y es simplemente que Y muere. No puede producir un peor resultado el que X forme su intención asesina, y entonces actúe con mucha maldad. Como afirmé, podemos concluir que la evitación de la maldad es un mero medio.

En casos así, muchos piensan que la maldad no tiene peores consecuencias. ¿Las tendría mejores el que hubiera más actos que fueran moralmente correctos, y más deberes que fueran cumplidos? Yo podría a menudo prometer hacer lo que de todos modos intentaba hacer. Así haría verdadero que más deberes fueran cumplidos. Pero nadie va a pensar que esto tendría un resultado mejor, o que fuese lo que debo hacer.

Supongamos a continuación que la pobreza queda abolida, que los desastres naturales dejan de ocurrir, que las personas dejan de sufrir enfermedades físicas y mentales, y que en muchas otras maneras las personas dejan de necesitar la ayuda de los demás. Todos estos cambios serían en un sentido buenos. ¿Serían malos en algún sentido? Es moralmente admirable ayudar a otros cuando se ven en dificultades, a costa de un considerable sacrificio para uno mismo. Pero en el mundo que he descrito muy pocos necesitan semejante ayuda. Habría muchos menos de estos actos moralmente admirables. ¿No sería malo esto? ¿No tendría un resultado en un respecto peor?

Si contestamos No, esto apoya una vez más la concepción de que actuar moralmente es un mero medio. Pero algunos de nosotros contestarían Sí. Pensaríamos que, en este respecto, el resultado *sería* peor. Y hay también muchas personas que mantienen una concepción diferente acerca de la maldad. Estas personas creerían que, comparada con la muerte accidental de Z, el que X asesine a Y sería, como consecuencia, mucho peor. ¿Puede aceptar un consecuencialista estas afirmaciones?

Esto depende de qué principios acepte. Consideremos en primer lugar a un utilitarista hedonista. Si el que X asesine a Y no causara más sufrimiento que la muerte accidental de Z, o una pérdida

de felicidad mayor, este utilitarista no podría afirmar que, como resultado, el asesinato es mucho peor. Volviendo a los actos moralmente admirables, todo lo que puede afirmar es lo siguiente. Una de las principales fuentes de felicidad es la creencia de que uno está ayudando a los demás de maneras relevantes. Por tanto sería malo en un sentido que muy pocas personas necesitasen tal ayuda.

Consideremos a continuación a un consecuencialista que aceptara la Teoría de la Lista Objetiva en relación con el propio interés. Según esta teoría, ser moral y actuar moralmente pueden ser en sí mismos buenos para nosotros, cualquiera que sean sus efectos. Pueden estar entre las cosas que son mejores para nosotros, o que hacen que nuestras vidas vayan mejor. Y ser moralmente malo puede ser, en sí mismo, una de las cosas peores para nosotros. Si un consecuencialista hiciera estas afirmaciones, podría negar que actuar moralmente y evitar la maldad son meros medios. Según cualquier versión plausible de C, es mejor que nuestras vidas vayan mejor. Según las afirmaciones recién hechas, actuar moralmente y evitar la maldad *son* partes del fin moral último dado a nosotros por C.

Según esta concepción, aunque estas forman parte de este fin, no son, como tales, fines últimos. Forman parte de este fin porque, igual que ser feliz, ser moral es una de las cosas que hacen que nuestra vida vaya mejor.

Un consecuencialista podría hacer una afirmación diferente. Podría afirmar que nuestro fin formal es, como tal, un fin sustantivo. Podría afirmar que sería peor si hubiera más maldad, aunque esto no fuera peor para nadie. De forma similar, podría ser mejor si más personas actuaran moralmente, aun cuando esto no fuese mejor para nadie. Un consecuencialista podría incluso afirmar que el logro de nuestro fin formal tiene absoluta prioridad sobre el logro de nuestros otros fines morales. Podría aceptar la concepción del Cardenal Newman. Newman pensaba que el dolor y el pecado eran malos los dos, pero que el pecado era infinitamente peor. Si toda la humanidad sufriera «el tormento más extremo», esto sería menos malo que si se cometiera un pecado venial [27].

[27] Newman, vol. I, p. 204.

Pocos consecuencialistas irían tan lejos. Pero como la concepción de Newman es una versión de C, no podemos afirmar que C da demasiado poco peso a la evitación de la maldad. C podría dar a este fin absoluta prioridad sobre todos nuestros otros fines morales.

Puede parecer de nuevo que C no es una teoría moral distintiva, sino que podría cubrir todas las teorías. Pero esto no es así. Los no consecuencialistas pueden afirmar, no que C da demasiado poca importancia a la evitación de la maldad, sino que C da esta importancia en el sentido equivocado. Según esta versión extrema de C, la evitación de la maldad sería uno de nuestros fines morales *comunes*. Un no-consecuencialista diría que yo no debo actuar mal, aun cuando a causa de mi acto hubiera mucha menos maldad por parte de otras personas. Según esta versión de C, yo no estaría actuando mal en este caso. Si hago lo que con más efectividad reduce la incidencia de la maldad, estoy haciendo lo que debo hacer.

20. Conclusiones

Ahora resumiré la segunda mitad de este capítulo. Asumí que, en los modos que he descrito, el Consecuencialismo es indirectamente contraproducente. Tendría peores consecuencias el que siempre estuviéramos dispuestos a hacer lo que tuviera mejores consecuencias. Si todos tuviéramos esta disposición, las consecuencias podrían ser mejores de lo que en realidad son, dado el modo en que la gente es realmente. Pero las consecuencias serían peores de lo que serían si tuviéramos otros determinados conjuntos de motivos causalmente posibles.

Pregunté si C falla en sus propios términos, cuando es indirectamente contraproducente. Si produjera un resultado peor el que estuviéramos siempre dispuestos a hacer lo que produjese un resultado mejor, C nos diría que no debiéramos tener esta disposición. Puesto que C hace esta afirmación, no falla en sus propios términos.

Supongamos que todos nosotros aceptáramos C. Nuestra teoría nos dice que deberíamos resolvernos a tener, o a mantener, uno de los mejores conjuntos de motivos posibles, en términos conse-

cuencialistas. Puesto que C es indirectamente contraproducente, esto implicaría lo siguiente. Si tenemos uno de los mejores conjuntos posibles de motivos, a veces actuaremos mal a sabiendas, según nuestra propia teoría. Pero, dada la razón especial por la que estamos actuando mal, no tenemos necesidad de considerarnos a nosotros mismos, cuando actuamos así, moralmente malos. Podemos pensar que estos son casos de *maldad inocente*. Podemos pensar esto porque estamos actuando sobre la base de un conjunto de motivos que sería incorrecto para nosotros resolvernos a perder.

Podrían venir implicadas algunas de estas afirmaciones aun cuando C no fuese indirectamente contraproducente. Vendrían implicadas si C fuera exigente de un modo no realista. Esto probablemente es cierto. Es probable que, aun cuando todos nosotros creyéramos en C, fuese causalmente imposible que llegásemos a estar dispuestos siempre a hacer lo que creemos que produciría un resultado mejor. Si esto es verdadero, C nos dice que tratemos de tener uno de los mejores conjuntos posibles de motivos.

Al hacer estas diversas afirmaciones, C es coherente. Y tampoco fracasa a la hora de tomarse la moralidad en serio. Aun si aceptamos estas afirmaciones, todavía habría muchos otros casos en que nos consideraríamos a nosotros mismos como moralmente malos. Esto sería así cuandoquiera que a sabiendas produjéramos un resultado *mucho* peor, y *no* lo hiciésemos porque tuviéramos uno de los mejores conjuntos posibles de motivos. Aunque no estuviéramos solamente interesados en la evitación de la maldad, puesto que también estaríamos interesados en tener los mejores motivos, todavía consideraríamos muchos actos como demostrando que el agente es moralmente malo.

Puede objetarse que estas afirmaciones asumen erróneamente el determinismo psicológico. Si un consecuencialista acepta esta objeción, tiene que matizar sus afirmaciones. Puede afirmar que, si tenemos uno de los mejores conjuntos de motivos, sería con frecuencia muy difícil para nosotros evitar hacer lo que creemos que es incorrecto. Tiene que admitir que, en estos casos, no somos *completamente* inocentes; pero somos malos sólo en un sentido muy débil.

Otra objeción es que una teoría moral aceptable no puede decirnos que nos determinemos a hacer lo que ella misma afirma que es incorrecto. Pero di un ejemplo en que esta objeción sería negada incluso por la mayoría de los que rechazan C.

Una tercera objeción es que, puesto que C es indirectamente contraproducente, no podemos en todos los casos evitar hacer lo que C afirma que es incorrecto. Puesto que no podemos hacer siempre lo que C afirma que debemos hacer, C exige lo imposible. Infringe la doctrina de que deber implica poder. Afirmé que esta objeción podría ser contestada.

Una cuarta objeción es que producirá un resultado mejor el que tengamos más creencias que se opongan a C. Si esto fuera verdadero, C sería modesta. Nos diría que creyésemos, no en sí misma, sino en alguna otra teoría. Puse en duda que esto fuese verdadero. Creo que C es, como mucho, parcialmente modesta y parcialmente esotérica. Podría tener mejores consecuencias que algunas personas creyeran en alguna otra teoría, pero no las tendría que nadie creyera en C. Y, aunque C fuese completamente modesta, creo que esto no pondría a C en tela de juicio. Si esto es así depende de cuál sea la mejor concepción de la moralidad y del razonamiento moral. Puesto que no he defendido ninguna de estas concepciones, no defendí plenamente mi creencia de que, si C fuera modesta, esto no pondría a C en tela de juicio.

Pregunté, por fin, si podemos aceptar la afirmación de C de que actuar moralmente es un mero medio. Si no podemos, C no necesita hacer esta afirmación. Podría incluso afirmar que la prevención de la maldad tiene absoluta prioridad sobre nuestros otros fines morales. Los *utilitaristas* no pueden hacer esta afirmación. Pero los *consecuencialistas* sí que pueden.

En la primera mitad de este capítulo discutí la teoría del Propio Interés sobre la racionalidad. En la segunda mitad, discutí el grupo de teorías morales que son consecuencialistas. Es plausible afirmar que todas estas teorías son indirectamente contraproducentes. Y podrían ser, tal vez, modestas. Pero, en el caso de estas teorías, ser indirectamente contraproducente no es ser irreparablemente contraproducente. Ni estos hechos proporcionan objeciones indepen-

dientes contra estas teorías. No muestran que estas teorías sean o falsas o indefendibles. Esto puede ser verdadero, pero los argumentos hasta aquí considerados no lo han demostrado. Como mucho demuestran que lo que se puede afirmar justificadamente es más complicado que lo que podemos haber esperado.

2

DILEMAS PRÁCTICOS

21. Por qué C no puede ser directamente contraproducente

He descrito el modo en que las teorías pueden ser indirectamente contraproducentes. ¿Cómo podrían ser *directamente* contraproducentes? Digamos que alguien *sigue con éxito la Teoría T* cuando tiene éxito al hacer el acto que, de los que son posibles para él, mejor consigue sus fines T-dados. Usamos *nosotros* para significar «los miembros de determinado grupo». Podríamos decir que T es

> *directa y colectivamente contraproducente* cuando es verdadero que, si *todos* nosotros seguimos con éxito T, causaremos con ello que nuestros fines T-dados sean peor conseguidos de lo que lo habrían sido si *ninguno* de nosotros hubiera seguido con éxito T.

Esta definición parece plausible. «Todos» y «ninguno» nos dan los casos más simples. Será suficiente discutir estos casos.

Aunque parece plausible, tenemos que rechazar esta definición. Deja de ser plausible cuando se aplica a ciertos *problemas de coordinación*. Se trata de casos en que el efecto del acto de cada persona

depende de lo que los otros hacen. Para un caso simple, ver página 167. Otro caso se muestra abajo.

		Tú haces (1)	Tú haces (2)	Tú haces (3)
Yo	hago (1)	El tercero mejor	El malo	El malo
Yo	hago (2)	El malo	El segundo mejor	El mejor
Yo	hago (3)	El malo	El mejor	El malo

Si los dos hacemos (1), los dos seguimos C con éxito. Como tú has hecho (1), yo habría obtenido un resultado peor si hubiera hecho (2) o (3). Y tú podrías decir lo mismo. Si en cambio los dos hacemos (2), ninguno ha seguido con éxito C. Puesto que tú has hecho (2), yo habría obtenido un mejor resultado si hubiera hecho (3). Y tú podrías decir lo mismo. Si los dos hacemos (1) antes que (2), los dos en vez de ninguno seguimos con éxito C. Pero con eso producimos un peor resultado, provocando que nuestro fin C-dado sea peor logrado. Según la definición dada arriba, C es aquí contraproducente.

Esta conclusión no está justificada. Es verdadero que, si los dos hacemos (1), los dos seguimos C con éxito. Pero si hubiéramos producido cualquiera de las mejores consecuencias, *también* habríamos seguido C con éxito. Si uno de nosotros hubiera hecho (2) y el otro hubiera hecho (3), cada uno habría hecho lo que, de los actos que eran posibles para él, hubiera producido un mejor resultado. La objeción a C aquí no es la de que es contraproducente. La objeción es que es *indeterminada*. Seguimos con éxito C *tanto* si los dos hacemos (1) *como* si uno de nosotros hace (2) y el otro hace (3). Como esto es verdadero, si ambos seguimos C con éxito, esto no *asegura* que nuestros actos produzcan conjuntamente uno de los mejores resultados posibles. Pero no asegura que *no* los produzcan. Si hubiésemos producido una de las mejores consecuencias, habríamos seguido C con éxito. C no nos *distancia* de las mejores consecuencias. La objeción no llega a tanto. C simplemente falla al dirigirnos a estas consecuencias. (Explicaré en la Sección 26 el modo en que esta objeción puede ser parcialmente anulada) [28].

Si C nos apartara de los mejores resultados, sería *cierto* que, si seguimos C con éxito, *no* produciremos uno de los mejores resultados. Esto sugiere otra definición. Llamemos a la teoría T

directa y colectivamente contraproducente cuando

(i) es *cierto* que, si todos nosotros seguimos T con éxito, con ello causaremos que nuestros fines T-dados sean peor logrados de lo que lo habrían sido si ninguno de nosotros hubiera seguido T con éxito, o bien

(ii) nuestros actos causarán que nuestros fines T-dados sean mejor logrados sólo si *no* seguimos con éxito T.

(ii) da expresión a la idea de que, para hacer que nuestros fines T-dados sean mejor logrados, tenemos que *desobedecer* a T. Por «cuando» no quiero decir «sólo cuando». No necesitamos incluir a todos los casos. Como expliqué, «nosotros» no significa «todo viviente», sino «todos los miembros de algún grupo».

¿Podrían ser verdaderas (i) y (ii) en el caso de C? ¿Podría ser verdadero que nosotros produzcamos un resultado mejor sólo si no seguimos con éxito C? ¿Podría ser cierto por tanto que, si todos antes que ninguno de nosotros seguimos C con éxito por ello produciremos un resultado peor? Ninguna de estas cosas es posible. Seguimos con éxito C cuando cada uno hace lo que, de todos los actos que son posibles para él, tiene las mejores consecuencias. Si nuestros actos conjuntamente producen el mejor resultado, todos nosotros tenemos que estar siguiendo con éxito C. No puede ser verdadero aquí de nadie que, si hubiera actuado en forma diferente, habría obtenido un resultado mejor. Según esta definición, C no puede ser directamente contraproducente.

C puede ser una teoría pluralista, que valore la bondad de los resultados apelando a varios principios diferentes. Uno sería alguna

[28] En estos parágrafos me limito a seguir a Regan.

tesis utilitarista acerca de perjuicios y beneficios; los otros podrían ser principios acerca de la distribución justa, o el engaño, o la coerción, o los derechos. Si estos y otros principios nos dicen que estemos de acuerdo en qué resultados serían mejores, el razonamiento recién dado será de aplicación. Tal teoría pluralista no puede ser directamente contraproducente, puesto que es *neutral con respecto al agente*: dando a todos los agentes sólo fines morales *comunes*. Si hacemos que estos fines comunes sean mejor conseguidos, tenemos que estar siguiendo con éxito esta teoría. Puesto que esto es así, no puede ser verdadero que hagamos que estos fines sean mejor conseguidos sólo si no seguimos esta teoría.

22. Cómo las teorías pueden ser directamente contraproducentes

¿Y qué ocurre si nuestra teoría es *relativa al agente*, dando a agentes *diferentes* fines? Puede que ahora seamos incapaces de aplicar la cláusula (ii) de mi definición. Si T da a diferentes personas fines diferentes, quizás no haya modo en que podamos lograr *mejor* los fines T-dados *de cada uno*. Pero podemos aplicar la cláusula (i), con una ligera revisión. Y daré otra definición. Llamaré a T

> *directa e individualmente contraproducente* cuando es cierto que, si alguien sigue T con éxito, hará con ello que sus propios fines T-dados sean peor logrados de lo que lo habrían sido si no hubiera seguido T con éxito,

y

> *directa y colectivamente contraproducente* cuando es cierto que, si todos nosotros seguimos con éxito T, con ello haremos que los fines T-dados *de cada uno* sean peor logrados de lo que lo habrían sido si ninguno de nosotros hubiera seguido T con éxito.

La teoría del Propio Interés da a diferentes agentes fines diferentes. ¿Podría esta teoría ser directa e individualmente contra-

producente? El fin que PI me da es que mi vida vaya, para mí, lo mejor posible. Yo sigo PI con éxito cuando hago lo que, de los actos que son posibles para mí, será mejor para mí. ¿Podría ser cierto que, si yo sigo PI con éxito, con ello produciré un resultado peor para mí? Esto no es posible. No es posible ni en el caso de un acto único ni en el de una serie de actos en tiempos diferentes. El argumento para esta segunda afirmación es como el argumento que di arriba. PI me da en tiempos diferentes uno y el mismo fin *común*: que mi vida vaya, para mí, lo mejor posible. Si mis actos en tiempos diferentes hacen que mi vida vaya lo mejor posible, tengo que estar siguiendo con éxito PI, al hacer cada uno de esos actos. Tengo que estar haciendo lo que, de los actos que son posibles para mí, sería lo mejor para mí. De modo que no puede ser cierto que, si siempre sigo PI con éxito, produzca con ello un resultado peor para mí.

Lo que puede ser peor para mí es estar siempre *dispuesto* a seguir PI. Pero en este caso no son mis actos los que son malos para mí, sino mi disposición. PI no puede ser directa e individualmente contraproducente. Puede ser sólo *indirecta* e individualmente contraproducente.

¿Pueden ser las teorías directa y *colectivamente* contraproducentes? Supongamos que la teoría T nos da a ti y a mí diferentes fines. Y supongamos que cada uno podría o bien (1) promover su propio fin T-dado o bien (2) promover el del otro de una manera más efectiva. Los resultados se muestran abajo.

		Tú haces (1)	Tú haces (2)
Yo	hago (1)	El fin T-dado de cada uno se logra el tercero mejor	El mío se logra el mejor, el tuyo el peor
Yo	hago (2)	El mío se logra el peor, el tuyo el mejor	El fin T-dado de cada uno se logra el segundo mejor

Supongamos finalmente que la elección de ninguno afectará a la del otro. Entonces será verdadero de cada uno que, si él hace (1) en vez de (2), con ello hará que su propio fin T-dado sea mejor logrado. Esto es así haga el otro lo que haga. Los dos seguimos con éxito T sólo

si los dos hacemos (1) en vez de (2). Sólo entonces cada uno está haciendo lo que, de los actos que son posibles para él, mejor logra su fin T-dado. Pero es cierto que si los dos antes que ninguno seguimos T con éxito —si los dos hacemos (1) en vez de (2)— con ello haremos que el fin T-dado de cada uno sea peor logrado. La teoría T es aquí directa y colectivamente contraproducente.

Casos como éste tienen gran importancia práctica. Los más simples pueden ocurrir cuando

(a) la teoría T es relativa al agente, y da a los diferentes agentes fines diferentes,
(b) el logro de los fines T-dados de cada persona depende parcialmente de lo que los otros hagan, y
(c) lo que cada uno hace no afectará a lo que esos otros hagan.

23. Los Dilemas del Prisionero y los bienes públicos

Estas tres condiciones con frecuencia se dan si T es la teoría del Propio Interés. PI es a menudo directa y colectivamente contraproducente. Estos casos tienen un nombre engañoso tomado de un ejemplo. El *Dilema del Prisionero*. A ti y a mí se nos interroga por separado acerca de algún crimen cometido en común. Los resultados se muestran abajo. Haga lo que haga el otro, será mejor para cada uno si confiesa. Si confiesa, cada uno puede tener por seguro que se ahorrará dos años de cárcel. Pero si los dos confiesan eso será peor para cada uno que si los dos guardan silencio.

		Tú	
		confiesas	guardas silencio
Yo	confieso	A cada uno le condenan a 10 años	Quedo libre, a ti te condenan a 12 años
	guardo silencio	Me condenan a 12 años, tú quedas libre	A cada uno le condenan a 2 años

Simplifiquemos. Será peor para ambos si cada uno antes que ninguno hace lo que será mejor para él. Un caso ocurre cuando

(*La Condición Positiva*) cada uno podría o bien (1) darse a sí mismo el menor de dos beneficios o bien (2) dar al otro el beneficio más grande,

y

(*La Condición Negativa*) la elección de ninguno sería de otros modos mejor o peor para cada uno.

Cuando la Condición Positiva se cumple, los resultados son como se muestra abajo.

		Tú	
		haces (1)	haces (2)
Yo	hago (1)	Cada uno consigue el beneficio menor	Yo consigo ambos beneficios, tú no consigues ninguno
	hago (2)	Yo no consigo ningún beneficio, tú consigues los dos	Cada uno consigue el beneficio mayor

Si añadimos la Condición Negativa, el diagrama se convierte en el que se muestra abajo.

		Tú	
		haces (1)	haces (2)
Yo	hago (1)	Lo tercero mejor para cada uno	Lo mejor para mí, lo peor para ti
	hago (2)	Lo peor para mí, el mejor para ti	Lo segundo mejor para los dos

Parte de la Condición Negativa no puede mostrarse en este diagrama. No tiene que haber *reciprocidad alguna*: tiene que ser verdadero que ninguna elección haría que el otro hiciera la misma elección. Será entonces mejor para cada uno que haga (1) en vez de (2). Esto es cierto haga lo que haga el otro. Pero si los dos hacen (1) esto será para cada uno peor que si los dos hacen (2).

¿Cuándo no podría la elección de ninguno afectar a la del otro? Sólo cuando cada uno tuviera que hacer una elección final antes de enterarse lo que el otro eligió. Fuera de las prisiones, o de los despachos de los *teóricos de los juegos*, esto se cumple raras veces. Ni tampoco aseguraría la Condición Negativa. Podría haber, por ejemplo, reciprocidad diferida. La elección de uno podría afectar a si resulta después perjudicado o beneficiado por el otro. Raramente somos capaces, por tanto, de saber que nos enfrentamos a un Dilema del Prisionero de Dos Personas.

Esta última afirmación viene apoyada por la extensa literatura sobre los Dilemas del Prisionero. Esta literatura describe pocos Casos convincentes de Dos Personas. Mi Condición Negativa casi nunca se da.

Uno de los casos más discutidos es el de la carrera de armamentos entre los Estados Unidos y la Unión Soviética. A menudo se dice que este es un Dilema del Prisionero. ¿Debería cada una de estas naciones desarrollar secretamente nuevas armas? Si sólo una lo hace, puede ser capaz más tarde de dar órdenes a la otra. Esto sería su mejor resultado y el peor de la otra. Si las dos lo hacen, permanecerán igual, pero a un gran coste, y con la inseguridad de una competición continua. Esto sería lo tercero mejor para ambas. Lo segundo mejor para ambas sería que ninguna desarrollara secretamente nuevas armas. Cada una debería desarrollar nuevas armas puesto que eso producirá su tercer mejor resultado antes que su peor resultado si la otra hace lo mismo, y su mejor resultado en vez de su segundo mejor resultado si la otra no hace lo mismo. Pero si las dos desarrollan nuevas armas esto será para las dos peor que si ninguna lo hiciera [29].

[29] Esto se defiende, por ejemplo, en Gauthier (1), y en Brams.

Aquí puede cumplirse parte de mi Condición Negativa. Si las nuevas armas pueden ser desarrolladas secretamente, cada nación tiene que hacer su elección antes de enterarse de lo que la otra eligió. Sobre la cuestión de la investigación, el razonamiento que se acaba de dar puede ser correcto. Pero es dudoso si se aplica a la producción o al despliegue de nuevas armas, donde cada una pueda saber lo que la otra está haciendo. Y tampoco está claro que el mero progreso en investigación pueda poner en condiciones a cada una de dar órdenes a la otra. Además, esta es una situación *repetida* o *continuada*. Decisiones similares tienen que tomarse una y otra vez. Por esto deja de estar claro que actuar de una de las dos maneras será con seguridad mejor para cada nación. La elección hecha por cada una puede afectar a las ulteriores elecciones hechas por la otra.

Una gran parte de la literatura discute este tipo de caso repetido: los que engañosamente se llaman *Dilemas del Prisionero Repetidos*. Se han hecho muchos experimentos para ver cómo actúan pares de personas en tales casos [30]. Aparte de tal trabajo experimental, ha tenido lugar mucha discusión teórica de los «Dilemas del Prisionero Repetidos». Aunque es de interés, esta discusión es aquí irrelevante. Deberíamos distinguir dos tipos de casos. En el primero, cada persona sabe que se enfrentará a un número determinado de «Dilemas del Prisionero Repetidos». Puesto que en la mayoría de los casos que tienen importancia práctica esto no ocurre, discutiré estos casos en la nota [31].

En la mayoría de los casos importantes, no sabemos cuántas veces nos enfrentaremos a los «Dilemas del Prisionero Repetidos».

[30] En estas pruebas, a las dos elecciones se las llama *cooperativa y no cooperativa*. Se ha demostrado, por ejemplo, que si los participantes toman 5 miligramos de Valium, aumenta la probabilidad de que cooperen. En relación con esto son de interés muchos números de *The Journal of Conflict Resolution*, Ann Arbor, Michigan.

[31] Se ha defendido con frecuencia que, si cada persona sabe que se enfrentará a un número determinado de «Dilemas del Prisionero Repetidos», será mejor para cada una hacer siempre la elección no cooperativa. Esta elección será mejor para cada una en la última vuelta del juego. Y como todas las personas lo saben, esta elección será mejor para cada una de ellas en la penúltima vuelta. Un razonamiento parecido se remonta hasta la primera vuelta.

Según mi definición, los que se enfrentan a tales series de casos no se enfrentan ni siquiera a un solo Dilema del Prisionero *verdadero*. No es verdadero, de tales personas, que sea peor para las dos si cada una en vez de ninguna hace lo que será mejor para ella. Esto no es verdadero porque, en estos «Dilemas del Prisionero Repetidos», ya no está claro cuál de las dos elecciones será mejor para uno mismo. Esto ocurre porque la elección de uno puede afectar a las posteriores elecciones hechas por el otro. Si uno hace la elección cooperativa, puede llevar al otro más tarde a hacer lo mismo. Como los teóricos de juegos dicen, si consideramos todas sus posibles consecuencias, ninguna elección es *dominante*, es mejor para uno mismo con certeza. La cuestión que suscitan tales casos es, por tanto, una cuestión *interna* para un teórico del Propio Interés. Si tu fin es hacer lo mejor que puedas hacer para ti mismo, ¿cómo deberías actuar en una serie de «Dilemas del Prisionero Repetidos»? En un verdadero Dilema del Prisionero, las cuestiones que surgen son muy diferentes. En un dilema verdadero, si uno actúa de una de las dos maneras, es *seguro* que esto será mejor para uno mismo, no sólo inmediatamente sino a largo plazo y consideradas las cosas en su conjunto. El problema planteado no es el problema interno de cómo puede uno perseguir de la mejor manera sus propios intereses. El problema es que, si cada uno antes que ninguno hace lo que es seguro que es mejor para él mismo, esto será peor para ambos.

Aunque raramente podamos saber que encaramos un Dilema del Prisionero de Dos Personas, con frecuencia podemos saber que nos enfrentamos a Versiones de Muchas Personas. Y estas tienen una gran importancia práctica. El raro Caso de Dos Personas es importante sólo como modelo para las Versiones de Muchas Personas. Nos enfrentamos a

> un *Dilema de Muchas Personas* cuando es seguro que, si cada uno antes que ninguno de nosotros hace lo que será mejor para él mismo, será peor para todos.

Esta definición incluye sólo a los casos más simples. Como antes, «todos» significa «todas las personas de un grupo».

Un Caso de Muchas Personas es el *Dilema del Samaritano*. Cada uno de nosotros podría ayudar de vez en cuando a un desconocido a un coste menor para sí mismo. Cada uno podría, aproximadamente con la misma frecuencia, ser ayudado de manera parecida. En comunidades pequeñas, el coste de ayudar podría ser indirectamente equilibrado. Si yo ayudo, esto puede hacer que yo, a cambio, sea ayudado más tarde. Pero en comunidades grandes no es probable. Aquí puede ser mejor para cada uno que nunca ayude. Pero será peor para cada uno que nadie ayude nunca. Cada uno podría salir ganando si no ayuda nunca, pero perdería, y perdería más, si no es nunca ayudado.

Muchos casos ocurren cuando

> (*Las Condiciones Positivas*) (i) cada uno de nosotros podría, a cierto coste para él mismo, dar a los otros una suma total mayor de beneficios, o de beneficios esperados; (ii) si cada uno antes que ninguno diera este mayor beneficio a los otros, cada uno recibiría un beneficio, o un beneficio esperado, mayor; y
>
> (*La Condición Negativa*) no habría efectos indirectos que cancelaran estos efectos directos.

Las Condiciones Positivas cubren muchos tipos de casos. En un extremo, cada uno podría dar a *uno* de los otros un beneficio mayor. Un ejemplo es el Dilema del Samaritano. En el otro extremo, cada uno podría dar a *todos* los demás una suma total de beneficios mayor. (En el segundo extremo, donde cada uno podría beneficiar a todos los otros, (ii) es redundante puesto que está implicada por (i). En los otros casos, (ii) es a menudo verdadera. Sería verdadera, por ejemplo, si los beneficios se extendieran al azar.)

Otra gama de casos implica las diferentes *probabilidades* de que lo que cada uno hace beneficiara a los demás. En un extremo, cada uno podría ciertamente dar a los otros una suma total de beneficios mayor. En el otro extremo, cada uno tendría una probabilidad muy pequeña de dar a los otros un beneficio mucho mayor. En esta gama de casos cada uno podría dar a los otros una suma mayor de beneficios *esperados*. Esta es el valor de los beneficios posibles multipli-

cado por la probabilidad de que el acto los produzca. Cuando los efectos de nuestros actos son inciertos, mi definición del dilema necesita revisión. En estos casos no es seguro que, si cada uno en vez de ninguno hace lo que será mejor para él, esto será peor para todos. Nos enfrentamos a

> un *Dilema Arriesgado* cuando es seguro que, si cada uno en vez de ninguno se da a sí mismo un beneficio esperado, esto o bien reducirá el beneficio esperado para todos, o bien impondrá a todos un perjuicio o coste esperado.

En algunos Casos de Muchas Personas, sólo se cumplen las Condiciones Positivas. En estos casos, como los números implicados son *lo suficientemente pequeños*, lo que cada uno hace podría afectar a lo que la mayoría de los demás hace. Estos casos son importantes desde el punto de vista práctico. Hay muchos que implican a naciones, o a corporaciones de negocios, o a sindicatos. Casos tales tienen algunos de los rasgos de un genuino Dilema del Prisionero. Pero carecen del rasgo central. Como el acto de cada uno puede afectar a los actos de bastantes de los demás, no está claro cuál acto iría a favor de los intereses de cada uno. El problema planteado por tales casos es otra cuestión interna para un teórico del Propio Interés. En un genuino Dilema del Prisionero, no hay inseguridad en lo que respecta a qué acto, vistas las cosas en su conjunto, dará al agente un beneficio o beneficio esperado mayor. Los problemas planteados por los verdaderos Dilemas son muy diferentes.

Los Dilemas de Muchas Personas son, ya lo he dicho, extremadamente comunes. Una razón es esta. En un Caso de Dos Personas, es improbable que se cumpla la Condición Negativa. Esto puede necesitar que se asegure especialmente, sea por oficiales de prisión o por teóricos de los juegos. Pero en casos que implican a muchas personas, la Condición Negativa se cumple naturalmente. No hace falta que ocurra que cada uno de nosotros tenga que actuar antes de enterarse de lo que los otros. hacen. Aunque esto no suceda, si somos muy numerosos, lo que cada uno hace sería muy improbable que afectara a lo que la mayor parte de los demás haga. Puede afectar a lo que unos cuantos hagan; pero esto casi nunca marcaría una diferencia suficiente.

Los verdaderos Dilemas más comunes son los *Dilemas del Contribuyente*. Son los que implican *bienes públicos*: consecuencias que benefician incluso a aquellos que no ayudan a producirlos. Puede ser verdadero de cada persona el que, si ayuda, contribuirá a la suma de beneficios, o de beneficios esperados. Pero sólo una porción muy pequeña del beneficio con el que contribuya volverá a *ella*. Puesto que su parte de aquello con lo que ha contribuido será muy pequeña, puede que no se le devuelva su contribución. De forma que tal vez sea mejor para cada uno no contribuir. Y esto puede ser así, sea lo que sea lo que los demás hagan. Pero será peor para cada uno si muy pocos contribuyen. Y si nadie contribuye esto será peor para cada uno que si todos lo hacen.

Muchos Dilemas del Contribuyente implican dos umbrales. En estos casos, hay dos números v y w tales que, si menos que v contribuyen, no se producirá ningún beneficio, y si más que w contribuyen, esto no incrementará el beneficio producido. En muchos de estos casos no sabemos lo que los otros están inclinados a hacer. Entonces no será seguro que si alguien contribuye beneficiará a los otros. Será verdadero sólo que dará a los otros un beneficio esperado. Un caso extremo es el de una votación, donde la brecha entre los dos umbrales puede ser la brecha de un único voto. El número w es aquí $v + 1$. Aunque una elección casi nunca es un genuino Dilema del Prisionero, valdrá la pena discutirla más adelante.

Algunos bienes públicos necesitan contribuciones financieras. Esto es lo que pasa con las carreteras, la policía o la defensa nacional. Otros necesitan esfuerzos cooperativos. Cuando en las grandes industrias los salarios dependen de los beneficios, y el trabajo es desagradable o una carga, puede ser mejor para cada uno que los otros trabajen más duro, y peor para cada uno que lo haga el mismo. Igual puede ocurrir con los campesinos en granjas colectivas. Un tercer tipo de bien público es la evitación de un mal. La contribución que se requiere aquí es a menudo el autocontrol. Tales casos pueden implicar

> *Personas que Viajan Cada Día de su Casa al Trabajo*: Cada uno va más rápido si él conduce, pero si todos conducen cada uno va más lento que si todos cogen el autobús;
>
> *Soldados*: Cada uno estará más seguro si se da la vuelta y sale corriendo, pero si todos lo hacen matarán a más que si ninguno lo hace;
>
> *Pescadores*: Cuando el mar está lleno de peces, puede ser mejor para cada uno si intenta coger más y peor para cada uno que todos lo hagan;
>
> *Campesinos*: Cuando hay exceso de población, puede ser mejor para cada uno si él o ella tiene más hijos, peor para cada uno si todos lo hacen [32].

Hay un sinnúmero de otros casos. Puede ser mejor para cada uno si contribuye a la polución, consume más energía, se salta las colas y rompe los acuerdos; pero si todos hiciéramos estas cosas, sería peor para cada uno que si ninguno las hiciera. Es muy a menudo verdad que, si cada uno antes que ninguno hace lo que es mejor para sí mismo, esto será peor para todos.

Estos Dilemas se describen usualmente en términos del propio interés. Como pocas personas se mueven puramente por el propio interés, esto puede parecer que reduce la importancia de estos casos. Pero en la mayoría de ellos ocurre lo siguiente. Si cada uno antes que ninguno hace lo que es mejor para sí mismo, *o para su familia, o para aquellos a los que quiere*, esto será peor para todos.

24. El problema práctico y sus soluciones

Supongamos que cada uno está dispuesto a hacer lo que es mejor para sí mismo, o para su familia, o para aquellos a quienes quiere.

[32] John Broome ha cuestionado esto con la observación de que «si un hijo extra produce más de lo que consume, su existencia es mejor para todos, y si consume más de lo que produce su existencia es peor para todos, incluidos sus propios padres». Creo que esta objeción queda contestada por la ventaja mucho mayor que reporta en la vejez tener más hijos.

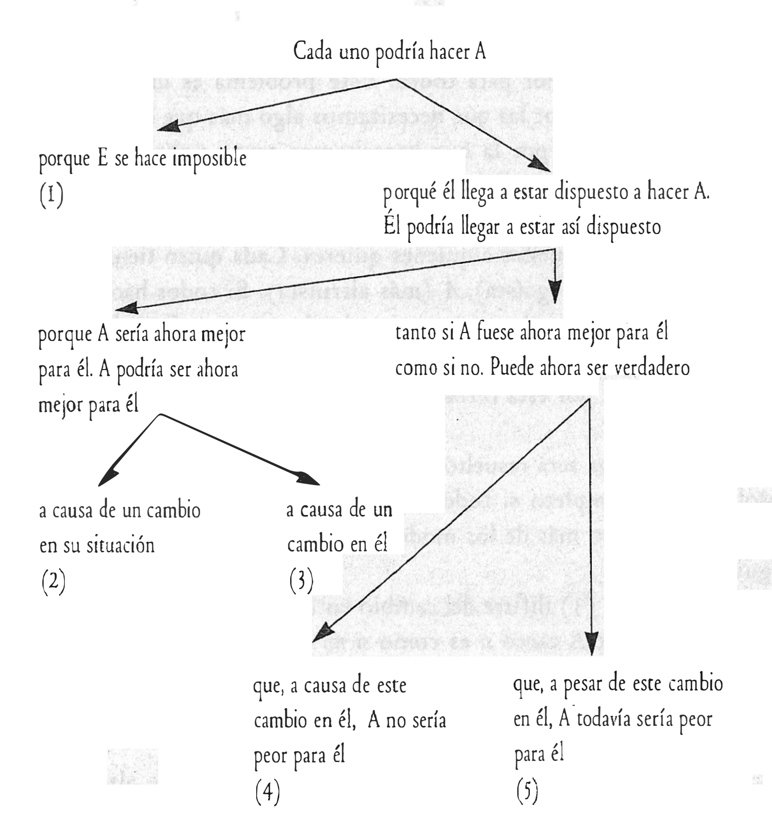
Se plantea entonces un *problema práctico*. Como no cambie algo, el resultado real será peor para todos. Este problema es una de las principales razones por las que necesitamos algo más que la economía del *laissez-faire* – por la que necesitamos tanto política como moral.

Usemos etiquetas. Y demos por entendidas las palabras «o para su familia, o para aquellos a quienes quiere». Cada quien tiene dos alternativas: *E* (más egoísta), *A* (más altruista). Si todos hacemos E eso será peor para cada uno que si todos hacemos A. Pero, hagan lo que hagan los demás, será mejor para cada quien si él hace E. El problema es que, por esta razón, cada quien está ahora dispuesto a hacer E.

Este problema será resuelto en parte si la mayoría hace A, y se resolverá por completo si todos lo hacen. Se puede alcanzar una solución en uno o más de los modos que se muestran en la página siguiente.

El cambio en (3) difiere del cambio en (4). En (4) alguien está dispuesto a hacer A tanto *si* es como *si no* es mejor para él. Es un simple efecto colateral el que, a causa de este cambio en él, A no sea peor para él. En (3) alguien está dispuesto a hacer A *sólo porque*, dado algún *otro* cambio en él, hacer A sería mejor para él.

Lo que va de (1) a (4) suprime el Dilema. La elección altruista deja de ser peor para cada uno. A menudo estas son buenas soluciones. Pero a veces son ineficientes o inalcanzables. Entonces necesitamos (5). (5) resuelve el problema *práctico*; pero no suprime el Dilema. Sigue habiendo un problema *teórico*. En este y en el próximo capítulo discuto cómo podemos resolver el problema práctico. Discuto el problema teórico en el capítulo 4.



En la solución (1), la elección que le beneficia a uno mismo se hace imposible. Esto es a veces lo mejor. En muchos Dilemas del Contribuyente, habría unos impuestos ineludibles. Pero con frecuencia sucede que (1) es una solución pobre. Las redes de pesca podrían ser destruidas, los soldados encadenados a sus puestos. Ambas situaciones tienen sus desventajas.

(2) es una solución menos directa. E sigue siendo posible, pero A se hace mejor para cada uno. Podría haber un sistema de recompensas. Pero, si el sistema funciona, todos tienen que ser recompensados. Puede ser mejor que la única recompensa sea evitar alguna penalización. Si esto funciona, nadie pagará. Si a todos los desertores se les pegara un tiro, no podría haber desertores.

La elección entre (1) y (2) es con frecuencia difícil. Consideremos el Dilema del Campesino, donde será mucho mejor para cada persona si él o ella tiene más hijos, y peor para cada persona si todos los tienen. Algunos países recompensan a las familias con dos niños. Ahora China recompensa a las familias que sólo tienen un niño. Pero donde el problema es más serio el país es demasiado pobre para dar a todos recompensas. Y, si tal sistema va a ser efectivo, las faltas de recompensa tienen que ser como castigos. Puesto que el sistema no sería completamente efectivo, algunos tendrían que soportar tales penalizaciones. Y éstas no sólo recaen sobre los padres sino también sobre los hijos.

Una alternativa es (1), donde se hace imposible que las personas tengan más de dos hijos. Esto conllevaría la esterilización obligatoria después del nacimiento de tu segundo hijo. Sería mejor que esta esterilización pudiera ser reversible, para el caso de que uno o los dos hijos murieran. Tal solución puede parecer horrenda. Pero podría recibir apoyo unánime en un referéndum. Sería mejor para todos los miembros de un grupo que ninguno antes que todos tuviera más de dos hijos. Si todos prefieren que sea esto lo que ocurra, todos pueden preferir y votar ese sistema de esterilización obligatoria. Y si fuera respaldado por unanimidad en un referéndum, esto eliminaría lo que es horrendo en la obligación. Y esta solución tiene ventajas si la comparamos con un sistema de recompensas y castigos. Como he dicho, cuando tal sistema no resultara completamente efectivo, los que tuvieran más hijos tendrían que pagar multas, como asimismo sus hijos. Puede que sea mejor que aquello por lo que se paga la multa sea, en vez de eso, hecho imposible [33].

(1) y (2) son soluciones políticas. Lo que se cambia es nuestra situación. De (3) a (5) son *psicológicas*. Somos nosotros los que cambiamos. Este cambio puede ser específico, resolviendo sólo un dilema. Los pescadores podrían volverse vagos, los soldados podrían llegar a preferir la muerte al deshonor, o ser instruidos en la obediencia automática. Aquí tenemos cuatro cambios de una clase más general:

[33] Cf. Sen y Runciman.

Podríamos volvernos *fiables*. Cada uno podría entonces prometer hacer A, a condición de que los demás hicieran la misma promesa.

Podríamos volvernos *reacios a ir de «polizones»*. Si cada uno cree que otros muchos harán A, él puede entonces preferir hacer su parte.

Podríamos volvernos *kantianos*. Entonces cada uno haría sólo lo que podría racionalmente querer que todos hicieran. Nadie podría racionalmente querer que todos hicieran E. Por consiguiente cada uno haría A.

Podríamos volvernos *más altruistas*. Dado un altruismo suficiente, cada uno haría A.

Estas son soluciones *morales*. Como podrían solucionar cualquier Dilema, son las soluciones psicológicas más importantes.

A menudo son mejores que las soluciones políticas. Esto se debe, en parte, a que no necesitan que se las imponga. Tomemos el Dilema del Samaritano. No puede hacerse imposible el no ayudar a los desconocidos. Los malos samaritanos no pueden ser cogidos y multados fácilmente. Los buenos samaritanos podrían ser recompensados. Pero para que esto se garantizara podría tener que intervenir la ley. Dados los costes administrativos, puede que esta solución no valga la pena. Sería mucho mejor si nos volviéramos directamente dispuestos a ayudar a los desconocidos.

No es suficiente saber cuál solución sería la mejor. Cualquier solución puede lograrse, o realizarse. Esto es a menudo más fácil con las soluciones políticas. Se pueden cambiar las situaciones con más facilidad que las personas. Pero a menudo nos enfrentamos a otro Dilema del Contribuyente, esta vez de segundo orden. Pocas soluciones políticas pueden conseguirse por parte de una persona sola. La mayoría requieren la cooperación de muchas personas. Pero una solución es un bien público, que beneficia a cada persona, haga o no haga su parte para llevarla a la práctica. En los grupos más grandes, será peor para cada uno si cumple con su parte. La diferencia que suponga su acción será demasiado pequeña como para recuperar su contribución.

Este problema puede ser pequeño en democracias bien organizadas. Puede bastar aquí con acometer el Dilema original ampliamente entendido. Esto puede ser difícil. Pero entonces podemos votar una solución política. Si nuestro gobierno responde a las encuestas, incluso no habría necesidad de plantear una votación.

El problema es mayor cuando no hay gobierno. Esto es lo que preocupaba a Hobbes. Ahora debería preocupar a las naciones. Un ejemplo es la proliferación de armas nucleares. Sin un gobierno mundial, puede ser difícil llegar a una solución [34].

El problema es enorme cuando a su solución se opone algún grupo dirigente. Este es el *Dilema de los Oprimidos*.

Tales Dilemas del Contribuyente a menudo demandan soluciones morales. A menudo necesitamos algunas personas que estén directamente dispuestas a hacer su parte. Si estas pueden cambiar la situación, para llegar a una solución política, esta solución puede recomendarse a sí misma. Pero sin que haya gente así, nunca podremos lograrla.

Las soluciones morales son por tanto con frecuencia las mejores; y a menudo son las únicas soluciones alcanzables. Por eso necesitamos los motivos morales. ¿Cómo podrían ser introducidos? Afortunadamente, ese no es nuestro problema. Existen. Así es como resolvemos muchos Dilemas del Prisionero. Lo que necesitamos es fortalecer estos motivos, y hacer que se extiendan más y más.

Con esta tarea la teoría ayuda. Los Dilemas del Prisionero necesitan ser explicados. También sus soluciones morales. Ambos han sido demasiado poco entendidos.

Una solución es, como vimos, un acuerdo condicional. Para que este sea posible, antes tiene que ocurrir que nos podamos comunicar todos. Si nos guiamos puramente por el propio interés, o nunca somos abnegados, la habilidad de comunicarnos raramente supondría una diferencia. En los grupos más grandes, sería irrelevante

[34] La versión del problema que tiene que ver con la proliferación nuclear es en un sentido menos grave, puesto que el número de los agentes es relativamente pequeño. Aquí es posible que lo que cada uno hace pueda afectar a lo que hacen los demás.

prometer que vamos a hacer la elección altruista, puesto que sería peor para cada persona si cumpliera su promesa. Pero supongamos que somos fiables. Cada uno puede ahora prometer hacer A, a condición de que *todos los demás* hagan la misma promesa. Si sabemos que todos somos fiables, cada uno tendrá un motivo para unirse a este acuerdo condicional. Cada uno sabrá que, como no se una, el acuerdo no tendrá efecto. Una vez que todos hayamos hecho esta promesa, todos haremos A.

Si somos muchos, la unanimidad será en la práctica difícil de conseguir. Si nuestro *único* motivo moral es la formalidad, entonces estaremos poco inclinados a lograr el acuerdo condicional conjunto. Sería probable que fuese peor para cada uno el unirse. (También estaremos poco inclinados siquiera a comunicarnos) [35].

[35] Supongamos que no exigimos tanto como la unanimidad. En una comunidad de 1.000 miembros, cada uno podría hacer que su promesa de hacer A esté condicionada a que se le unan al menos otros 900. No lograremos *esta* solución si nuestro *único* motivo moral es la formalidad. Probablemente sería peor para cada persona hacer esta promesa. Hay tres posibilidades. El número de los otros que se unen podría resultar que es (a) menos de 900, (b) 900, o (c) más de 900. Si ocurre (a), no se convierte en vinculante la promesa de nadie, de modo que podemos ignorar esta posibilidad al decidir lo que va a favor de nuestros intereses. Si ocurre (b), será mejor para cada persona que haga su promesa. De nuevo será verdadero de cada uno que, sin su promesa, todo el esquema fracasaría —no llegaría a ser vinculante la promesa de nadie más—. Eso sería peor para cada persona. Pero si lo que ocurre es (c), será peor para cada persona hacer su promesa. Será verdadero de cada uno que está comprometido ahora con la elección altruista. Como es digno de confianza, hará esta elección. Pero aun sin su promesa, las promesas de todos los demás habrían llegado a ser vinculantes. De modo que cada uno hará la elección que será peor para sí mismo, y no ganará nada a cambio. Lo que cada uno gana de las promesas de los demás lo habría ganado aunque no hubiera hecho él mismo la promesa. Fijémonos además en que, comparado con (b), es mucho más probable que ocurra (c). Es mucho más probable que más de 900 de los otros se unan que el que el número de los otros que se unen sea exactamente de 900. Así que, si cada uno hace esta promesa, es probable que esto vaya a ser peor para él. Si los números son mayores, digamos del orden de millones, esto sería aún más probable. O la promesa de cada persona no representa diferencia alguna, puesto que se unen demasiado pocos de los otros, o será peor para ella, puesto que más de los otros de lo que sería suficiente se unen. El único caso en que hacer la promesa va a resultar mejor para cada persona, ese en el que se une exactamente el número

Hay pocas personas cuyo único motivo moral sea la formalidad. Supongamos que también somos reacios a ir de «polizones». Si cada uno de nosotros tuviera este motivo, no desearía quedar al margen del acuerdo condicional conjunto. Preferirá unirse, aunque hacerlo vaya a ser peor para él. Esto resuelve el problema recién mencionado para el acuerdo conjunto. Y si bastantes personas son reacias a ir de «polizones», no habrá necesidad de un acuerdo real. Todo lo que se necesitaría es la seguridad de que habrá muchos que hagan A. Cada cual entonces preferiría hacer su parte. Pero una renuencia a ser «gorrones» no puede por sí misma crear esta garantía. De forma que hay muchos casos en que no nos da solución [36].

El Test Kantiano siempre podría proporcionarla. Este Test tiene sus propios problemas. ¿Podría yo querer racionalmente o que nadie practicara medicina o que todos la practicaran? Si refinamos el Test, podemos ser capaces de resolver muchos problemas. Pero en los Dilemas del Prisionero no se plantean. Estos son los casos en que naturalmente decimos, «¿Y qué si todo el mundo lo hiciera?» [37].

La cuarta solución moral es altruismo suficiente. No me refiero aquí al *puro* altruismo. Los altruistas puros, que no asignan ningún peso a sus propios intereses, pueden enfrentarse a análogos del Dilema del Prisionero. Puede ocurrir que, si todos en vez de ninguno hacen lo que es seguro que será mejor para los demás, esto será peor para todos [38]. Por «altruismo suficiente» me refiero a

especificado, será el más improbable. Puede haber algún modo de evitar este problema, quizás con una serie de «votos de tanteo» convergentes. Pero en casos en que intervienen grandes números hay otro problema. Costaría cierto esfuerzo poner a todos en condiciones de comunicarse, para llegar entonces a un acuerdo conjunto. Pero el acuerdo es un bien público, que beneficia a cada una de las personas tanto si ayuda como si no ayuda a producirlo. La simple formalidad no proporciona ninguna solución a este Dilema del Contribuyente.

[36] Véase Sen (3).

[37] Véase, por ejemplo, Strang, y Ullmann-Margalit.

[38] En un Caso de Dos Personas, supongamos que cada altruista puro podría o bien (1) darse a sí mismo un beneficio mayor, o bien (2) dar al otro un beneficio más pequeño. Aquí el altruismo puro sería peor para los dos. (La diferencia entre los beneficios puede tener que ser grande, ya que lo que va a favor de nuestros intereses depende en parte de cuáles son nuestros motivos.)

una preocupación suficiente por los demás, donde el caso límite es la benevolencia imparcial: un interés igual por todos, incluido uno mismo.

La cuarta solución es la menos comprendida. A menudo se afirma que, en los Dilemas del Contribuyente que involucran a muchísimas personas, lo que cada persona hace no representaría ninguna diferencia. Esto se ha pensado para demostrar que los altruistas racionales no contribuirían. Como mantendré en el próximo capítulo, esto no es así.

3

CUATRO ERRORES EN MATEMÁTICAS MORALES

A menudo se afirma que, en casos que involucran a muchísimas personas, ninguna elección altruista singular significaría una auténtica diferencia. Algunos de los que afirman esto creen que socava sólo la cuarta solución moral, la que nos es proporcionada por el altruismo suficiente. Mantienen que, en tales casos, como no *podemos* apelar a las *consecuencias* de nuestros actos, tenemos que apelar en vez de ello o al Test Kantiano, «¿Y qué si todo el mundo lo hiciera?», o a la renuencia a ser «gorrones» [39]. Pero si mi contribución me acarrea un coste real, y ciertamente no va a significar diferencia ninguna —no va a dar beneficios a los demás— puede que no me motive mi renuencia a ser «gorrón». Esta resistencia puede ser de aplicación sólo cuando creo que estoy obteniendo beneficios a costa de los demás. Si mi contribución no va a marcar ninguna diferencia, mi fallo en contribuir no será peor para los demás, de modo que no me estaré beneficiando a su costa. Puedo creer que el caso es como esos en que algún umbral ha sido claramente cruzado, de forma que cual-

[39] Véase, por ejemplo, Ewing, Miller y Sartorius, Meehl, o el muy influyente Olson, p. 64, y p. 159.

quier acto altruista ulterior es un puro gasto de esfuerzo. Esta creencia puede también socavar la solución kantiana. Si mi contribución no va a significar ninguna diferencia, puedo racionalmente querer que todos los demás hagan lo que yo hago. Puedo querer racionalmente que nadie contribuya si sabe que su contribución no significaría diferencia ninguna. Puesto que otros pueden pensar como yo, es de gran importancia que, en tales casos, podamos explicar por qué deberíamos contribuir apelando a las consecuencias de nuestros actos.

25. La Concepción de la Parte-del-Total

Podemos. Pero para hacerlo tenemos que evitar varios errores en matemáticas morales. Consideremos

> *La Primera Misión de Rescate*: Sé todo esto: cien mineros están atrapados en un pozo en donde hay una crecida de agua. A estos hombres se les puede sacar a la superficie en un ascensor levantado por unos pesos colocados sobre grandes palancas. Si otras tres personas y yo nos ponemos sobre una plataforma, esto proporcionará justo el peso suficiente para levantar el ascensor, y así salvaremos la vida de estos cien hombres. Si no me uno a esta misión de rescate, puedo irme a otro lugar y salvar, sin ayuda, la vida de otras diez personas. Hay un quinto rescatador potencial. Si yo me voy a la otra parte, esta persona se unirá a las otras tres, y entre las cuatro salvarán a los cien mineros.

Cuando yo pudiera actuar de varias maneras, ¿cómo debería decidir cuál acto iba a beneficiar a la gente en mayor medida? Supongamos en primer lugar que los cinco juntos vamos a salvar a los mineros. Según la *Concepción de la Parte-del-Total*, cada persona produce su parte del beneficio total. Como entre los cinco salvaríamos cien vidas, cada uno salva veinte vidas. De manera menos literal, el bien que cada uno hace es equivalente a la salvación de esta cantidad de vidas. Según esta concepción, debo juntarme con los otros cuatro, y salvar el equivalente de veinte vidas. No debería irme a otra

parte a salvar a las otras diez personas, puesto que entonces estaría salvando a menos personas. Pero esta es sin duda la respuesta equivocada. Si me voy con los otros cuatro, diez personas mueren innecesariamente. Como los otros cuatro, sin mi ayuda, salvarían a los cien mineros, yo debería ir a salvar a estas diez personas.

La Concepción de la Parte-del-Total podría ser revisada. Podría afirmarse que, cuando me uno a los otros que hacen el bien, el bien que yo hago no es simplemente mi parte del beneficio total producido. Debería restar de mi parte cualquier reducción que cause mi unirme al grupo en las partes de los beneficios producidos por los otros. Si yo me uno a esta misión de rescate, seré una de las cinco personas que juntas salvan cien vidas. Mi parte será veinte vidas. Si yo no me hubiera unido al grupo, los otros cuatro habrían salvado a los cien, y la parte salvada por cada uno habría sido de veinticinco vidas, o cinco más que si yo me uno al grupo. Al unirme al grupo reduzco las partes de los otros cuatro en un total de cuatro por cinco, o veinte vidas. Según la concepción revisada, mi parte del beneficio es por tanto veinte menos veinte, o cero. Por eso debería irme a salvar a las otras diez personas. La concepción revisada da la respuesta correcta.

Consideremos a continuación

> *La Segunda Misión de Rescate.* Como antes, la vida de cien personas se halla en peligro. Estas personas pueden ser salvadas si yo y otras tres personas nos juntamos en una misión de rescate. Nosotros cuatro somos los únicos que podríamos juntarnos para esta misión. Si cualquiera de nosotros falla, las cien personas morirán. Si yo fallo y no me uno al grupo, podría irme a otro sitio y salvar, sin ayuda, otras cincuenta vidas.

Según la Concepción Revisada de la Parte-del-Total, debo ir al otro sitio a salvar a estos otros cincuenta. Si en vez de eso me uno a esta misión de rescate, la parte del beneficio que produzco es sólo el equivalente de salvar veinticinco vidas. Por eso puedo hacer más bien si me voy a la otra parte y salvo a los cincuenta. Esto es claramente falso, desde el momento en que si actúo así se perderán cincuenta vidas más. Tengo que unirme *a esta* misión de rescate.

Tenemos que hacer una revisión ulterior. Tengo que añadir a mi parte del beneficio producido cualquier incremento que yo cause a las partes producidas por los otros. Si yo me uno, permito a cada una de las tres personas salvar, conmigo, cien vidas. Si no me uno al grupo, estos tres no salvarán ninguna vida. Mi parte es de veinticinco vidas, e incremento en setenta y cinco las partes producidas por los otros. Según esta concepción dos veces revisada, mi parte total es de cien vidas. Esta es también la parte total producida por cada uno de los otros. Puesto que cada uno cuenta como produciendo el *todo* de este beneficio total, esto no es una versión de la Concepción de la *Parte*-del-Total. Es una concepción muy diferente. Esta concepción dos veces revisada da la respuesta correcta en este caso. No es objeción contra esta concepción que afirme que cada uno salva cien vidas. Esto es lo que cada uno hace, no por él mismo, sino con la ayuda de los otros [40].

Esta concepción puede formularse con más simplicidad. Yo debería actuar de un modo cuya consecuencia fuese que sean salvadas la mayor cantidad de vidas. Más en general,

> (C6) Un acto beneficia a alguien si su consecuencia es que alguien es más beneficiado. Un acto perjudica a alguien si su consecuencia es que alguien es más perjudicado. El acto que beneficia a la gente en mayor medida es el acto cuya consecuencia es que la gente es beneficiada en mayor medida.

Estas afirmaciones implican, correctamente, que yo no debería unirme a la Primera Misión de Rescate sino que debería unirme a la Segunda.

Los consecuencialistas deberían apelar a (C6). También deberían otros, si asignan algún peso a lo que Ross denominó *Principio de Beneficencia*. Según cualquier teoría moral plausible, deberíamos a veces tratar de hacer lo que beneficiara a la gente en mayor medida.

(C6) podría necesitar una explicación ulterior. Supongamos que puedo hacer o bien (1) o bien (2). Al decidir cuál iba a beneficiar más a la gente, debería comparar *todos* los beneficios y las pérdidas que la gente recibiría más tarde si yo hiciese (1), y *todos* los beneficios y las pérdidas que la gente recibiría más tarde si yo hiciese (2). El acto que beneficia más a la gente es aquel que, en esta comparación, sería seguido por la mayor suma *neta* de beneficios – la mayor suma de beneficios menos pérdidas. Es irrelevante si, como a menudo ocurre, los actos de muchas otras personas fuesen también parte de la causa del recibir estos beneficios y pérdidas.

[40] Debo estos puntos a un incisivo comentario de Martin Hollis.

(C6) revisa el uso ordinario de las palabras «beneficio» y «perjuicio». Cuando afirmo haber beneficiado a alguien, se entiende usualmente que lo que quiero decir es que algún acto mío fue la causa principal o inmediata de algún beneficio recibido por esta persona. Según (C6), yo beneficio a alguien aunque mi acto sea una parte remota de la causa del recibir este beneficio. Todo lo que hace falta que ocurra es que, si yo hubiese actuado de otra forma, esta persona no habría recibido este beneficio. Afirmaciones similares se aplican a «perjuicio».

De un segundo modo revisa (C6) nuestro uso de «beneficio» y «perjuicio». Según el uso ordinario, a veces beneficio a alguien aunque lo que yo estoy haciendo no sea mejor para esta persona. Esto puede ocurrir cuando mi acto, aunque suficiente para producir algún beneficio, no es necesario. Supongamos que podría fácilmente salvar o la vida de J o el brazo de K. Sé que, si no salvo la vida de J, alguien lo hará con seguridad; pero nadie más puede salvar el brazo de K. Según nuestro uso ordinario, si yo salvo la vida de J, le hago un beneficio, y le hago un mayor beneficio que el que yo haría a K si le salvara su brazo. Pero, para propósitos morales, este no es el modo de medir beneficios. Al tomar mi decisión, debería ignorar este beneficio a J, como (C6) me dice que haga. Según (C6), yo *no* beneficio a J cuando salvo su vida. No es verdadero que la consecuencia de mi acto es que J es beneficiado más. Si yo hubiera actuado de diferente modo, algún otro habría salvado la vida de J. (C6) implica correctamente que debo salvarle el brazo a K. *Este* es el acto cuya consecuencia es que la persona se beneficia más. Según el uso revisado de «beneficio», este es el acto que beneficia a la persona más.

El Primer Error en matemáticas morales es la Concepción de la Parte-del-Total. Deberíamos rechazar esta concepción, y apelar en su lugar a (C6)

26. Ignorar los efectos de conjuntos de actos

Es natural asumir

> (El segundo error). Si algún acto es correcto o incorrecto *a causa de sus efectos*, los únicos efectos relevantes son los efectos de este acto particular.

Esta asunción está equivocada al menos en dos clases de casos.

En algunos casos, los efectos están *sobredeterminados*. Consideremos

> *Caso Uno*. X e Y disparan simultáneamente y me matan. Cada disparo, por sí solo, me habría matado.

Ni X ni Y obran de un modo cuya consecuencia sea que una persona extra muera. Dado lo que el otro hace, es verdadero de cada uno que, si no me hubiera disparado, esto no habría representado ninguna diferencia. De acuerdo con (C6), ni X ni Y me perjudican. Supongamos que cometemos el Segundo Error: asumimos que, si un acto es incorrecto por razón de sus efectos, los únicos efectos relevantes son los efectos de este acto particular. Puesto que ni X ni Y me perjudican, nos vemos forzados a la absurda conclusión de que estos dos asesinos no obran mal.

Alguien podría coger este caso para demostrar que deberíamos rechazar (C6). Hay una alternativa mejor. Deberíamos añadir

> (C7) Aunque un acto no perjudique a nadie, este acto puede ser incorrecto porque es uno de un *conjunto* de actos que, *juntos*, perjudican a otras personas. De forma parecida, aunque determinado acto no beneficie a nadie, puede ser lo que alguien deba hacer, porque sea uno de un conjunto de actos que, juntos, benefician a otras personas.

X e Y obran mal porque *entre los dos* me perjudican. Entre los dos me matan. (C7) debería aceptarse incluso por los no consecuencialistas. Según cualquier teoría moral plausible, es un error en este tipo de casos considerar solamente los efectos de actos singulares. Según cualquier teoría plausible, aunque cada uno de nosotros no haga daño a nadie, podemos estar obrando mal si entre todos hacemos daño a otras personas.

En el Caso Uno, los actos sobredeterminantes son simultáneos. ¿Qué deberíamos afirmar en casos en que esto no es así? Consideremos

> *Caso Dos*. X me engaña para que yo beba veneno, un veneno que causa una muerte dolorosa en unos cuantos minutos. Antes de que este veneno haga el más mínimo efecto, Y me mata sin dolor.

Aunque Y me mata, el acto de Y no es peor para mí. (C6) implica por tanto que, al matarme, Y no me perjudica. (El acto de Y es de algún modo levemente peor para mí, puesto que acorta mi vida en unos pocos minutos. Esto es compensado por el hecho de que Y me salva de una muerte penosa.) (C6) también implica que X no me perjudica. Como en el Caso Uno, ni X ni Y me perjudican. (C7) implica correctamente que X e Y obran mal porque entre los dos me perjudican. Entre los dos me perjudican porque, si *ambos* hubieran actuado de forma diferente, yo no habría muerto.

Aunque (C7) da la respuesta correcta aquí, puede parecer que este caso proporciona una objeción contra (C6). Puede parecer absurdo afirmar que, matándome, Y no me está perjudicando. Pero consideremos

> *Caso Tres*: Como antes, X me engaña para que beba un veneno de una clase que causa una muerte dolorosa en unos pocos minutos. Y ahora sucede que Y sabe que puede salvar *tu* vida si actúa de un modo cuyo efecto colateral inevitable es mi muerte inmediata e indolora. Como Y también sabe que yo estoy a punto de morir con dolor, procede de esa manera.

(C6) implica que Y debe actuar de ese modo, puesto que él no me va a perjudicar, y en cambio te beneficiará enormemente a ti.

Esta es la conclusión correcta. Como el acto de Y no es peor para mí, es moralmente irrelevante que Y me mate. Es también moralmente irrelevante que X *no* me mate. (C6) implica correctamente que X obra mal. Aunque X no me mata según el uso ordinario de «matar», él es aquí el verdadero asesino. X me hace daño, y obra mal, porque es cierto que, si él no me hubiera envenenado, Y no me habría matado. Si X hubiera actuado de otra manera, yo *no* habría muerto. Y, por su parte, no me hace daño, porque, si *Y* hubiera actuado de modo diferente, esto no habría representado ninguna diferencia en la cuestión de si yo hubiera muerto. Como Y no me perjudica, y él te beneficia a ti enormemente, Y hace lo que debe hacer.

Como muestran estas observaciones, el Caso Dos no proporciona ninguna objeción contra (C6). En el Caso Tres, (C6) implica de manera correcta que Y debe obrar como lo hace, porque no me perjudica. En el Caso Dos, el acto de Y me afecta justamente de la misma manera. Tuve razón por tanto al afirmar que, en el Caso Dos, Y no me perjudica. Y obra mal en el Caso Dos porque él forma parte intencionalmente de un grupo que en unión me perjudica.

Puede objetarse que, si *esta* es la razón por la cual, en el Caso Dos, Y obra mal, Y tiene que obrar mal en el Caso Tres. Puede pensarse que, aquí también, Y forma parte intencionalmente de un grupo que en unión me perjudica.

Esta objeción muestra la necesidad de otra afirmación. En el Caso Tres es verdadero que, si los dos X e Y hubieran actuado de forma diferente, yo no habría sido perjudicado. Pero esto no muestra que X e Y juntos me perjudiquen. Es también verdadero que, si X, Y *y Fred Astaire* hubieran todos actuado de manera diferente, yo no habría sido perjudicado. Pero esto no convierte a Fred Astaire en miembro de un grupo que de forma conjunta me perjudica. Deberíamos afirmar

> (C8) Cuando determinado grupo, de manera conjunta, perjudica o beneficia a otras personas, este grupo es el grupo *más pequeño* del que es verdadero que, si todos hubieran actuado de forma diferente, esas otras personas no habrían sido perjudicadas, o beneficiadas.

En el Caso Tres, este «grupo» consta de X. Es verdadero de X el que, si él hubiera actuado de manera diferente, Y lo había hecho así también, y yo no habría sido perjudicado. Ninguna afirmación así es verdadera de Y. En el Caso Dos no es verdadero ni de X ni de Y el que, si uno de ellos hubiera actuado de manera diferente, yo no habría sido perjudicado. Yo no habría sido perjudicado sólo si *ambos* hubieran actuado de manera diferente. Tampoco habría sido yo perjudicado si X, Y y Fred Astaire hubieran actuado de forma diferente. Pero (C8) implica correctamente que Fred Astaire no es miembro del grupo que de manera conjunta me perjudica. Este grupo consta de X e Y.

Consideremos a continuación

> *La Tercera Misión de Rescate*. Como antes, si cuatro personas se colocan sobre una plataforma, se salvarán las vidas de cien mineros. Cinco personas se colocan sobre la plataforma.

Dado lo que los otros hacen, es verdadero de cada una de estas cinco personas que su acto no representa ninguna diferencia. Si uno no se hubiera subido a la plataforma, los otros cuatro habrían salvado a los cien mineros. Aunque ninguno, por sí mismo, representa diferencia alguna, los cinco juntos salvan a los cien mineros. Este caso muestra la necesidad de añadir alguna afirmación ulterior a (C8). En este caso no hay *un* grupo que sea el más pequeño, que en conjunto salve las cien vidas. Volveré a tales casos en la Sección 30.

Hay un segundo tipo de casos en que deberíamos considerar los efectos de conjuntos de actos. Estos son los *problemas de coordinación*. Un ejemplo se muestra abajo.

		Tú: haces (1)	Tú: haces (2)
Yo	hago (1)	El segundo mejor	El malo
	hago (2)	El malo	El mejor

Supongamos que aplicamos el Consecuencialismo sólo a actos singulares. Entonces afirmaremos que cada persona ha seguido con éxito C si ha llevado a cabo el acto, de los que son posibles para ella, cuya consecuencia es el resultado mejor. Como antes vimos, en los problemas de coordinación, C será entonces *indeterminado*. En este caso seguimos con éxito C *tanto* si ambos hacemos (2) *como* si ambos hacemos (1). Supongamos que ambos hacemos (1). Dado lo que tú has hecho, yo he llevado a cabo el acto cuya consecuencia es la mejor. El resultado habría sido peor si yo hubiera hecho (2). Las mismas afirmaciones se aplican a ti. Si ambos hacemos (1) los dos seguimos con éxito C, pero no hemos producido el resultado mejor posible.

Los consecuencialistas deberían afirmar

> (C9) Supongamos que cada uno de nosotros ha producido la mejor consecuencia posible, en vista de lo que los otros hicieron. Cada uno, entonces, ha actuado bien. Pero nosotros juntos podemos haber actuado mal. Esto será así si nosotros juntos hubiéramos podido producir una consecuencia mejor.

Esta es una afirmación acerca de la corrección *objetiva*, o lo que de hecho producirá el mejor resultado. Si C incluye esta afirmación, deja de ser indeterminado en esta clase de casos. (Cuando decidimos qué hacer, deberíamos preguntar en cambio qué es *subjetivamente* correcto, o qué tendrá más probabilidad, dadas nuestras creencias, de tener la mejor consecuencia. En la mayoría de los problemas de coordinación de esta clase, es subjetivamente correcto para cada uno de nosotros ponerse como objetivo el mejor resultado, puesto que esto es lo que probablemente harán los otros) [41].

[41] (C9) difiere del más complicado *Consecuencialismo Cooperativo* presentado en Regan. Más que revisar la afirmación del consecuencialista del acto acerca de cuándo *cada uno* de nosotros obra correctamente, (C9) se limita a añadir una afirmación acerca de cuándo *nosotros* obramos correctamente. (Estas afirmaciones pueden parecer inconsistentes. Si cada uno de nosotros ha obrado correctamente, ¿cómo podemos nosotros juntos haber actuado incorrectamente? Pero no hay conflicto aquí. Para los consecuencialistas, un acto es objetivamente correcto si tiene

27. Ignorar las pequeñas probabilidades

Volvamos ahora a esos Dilemas del Prisionero que involucran a muchísimas personas. Se afirma con frecuencia que, en estos casos, no podemos apelar a las consecuencias de nuestros actos. Esto implica como mínimo dos errores más.

Uno tiene que ver con esos casos en que cada acto altruista tiene una probabilidad extremadamente pequeña de producir algún beneficio extremadamente grande. A veces se afirma que, por debajo de un determinado umbral, las probabilidades extremadamente pequeñas no tienen ningún significado racional ni moral.

Este error a menudo se comete en las discusiones sobre elecciones. Aunque una elección no sea un Dilema del Prisionero puro, puede ilustrar este error. Muchos autores afirman que, en unas elecciones a escala nacional, no se puede justificar el hecho de que uno vote, simplemente apelando a las consecuencias del propio acto [42]. Pero esto es con frecuencia falso. Supongamos que, si yo voto, esto llevará consigo ciertos costes, y no reportará beneficios aparte del posible efecto sobre quién gana las elecciones. Partiendo de estas asunciones, el hecho de que yo vote no puede justificarse en términos del propio interés. Pero a menudo puede justificarse en términos consecuencialistas. Cuando no puedo predecir los efectos de mi acto, C me dice que haga lo que produzca el mayor beneficio *esperado*. El beneficio esperado de mi acto es el beneficio posible multiplicado por la probabilidad de que mi acto lo produzca. Quizás pueda justificar el hecho de que voto apelando a este beneficio.

Consideremos unas elecciones presidenciales en los Estados Unidos. Si yo voto, puede haber una probabilidad muy pequeña de que mi voto signifique una diferencia. Según una estimación, si emito mi voto en uno de los grandes estados marginales, de los que

los mejores efectos posibles, dadas las circunstancias. Cuando preguntamos si cada uno obró correctamente, las circunstancias incluyen lo que hicieron los otros. Cuando preguntamos si nosotros obramos correctamente, esto no es así.)

[42] Véase Meehl.

podría salir cualquier cosa, la diferencia que esto significará será alrededor de uno entre cien millones. (La estimación es difícil. No se debería asumir que cualquier pauta de votos es igual de probable que cualquier otra. Pero algunos autores están de acuerdo en que esta probabilidad es alrededor de una entre cien millones) [43].

Llamemos a los dos candidatos Superior e Inferior. Y supongamos que, si el próximo presidente es Superior, esto beneficiará al norteamericano medio. Habrá algunos norteamericanos que saldrán perdiendo. Habría sido mejor para estos norteamericanos que hubiera ganado Inferior. Pero las pérdidas de estos norteamericanos —la minoría rica— serán compensadas por los beneficios a todos los demás norteamericanos. Esta es la razón por la que Superior es el mejor candidato. Si él es elegido, esto producirá una mayor suma neta total de beneficios menos cargas. El beneficio neto medio para los norteamericanos es esta suma total dividida por el número de norteamericanos. Por simplicidad, voy a ignorar los efectos sobre los no norteamericanos. Si mi voto tiene una probabilidad de uno entre cien millones de afectar al resultado, el beneficio *esperado* del hecho de que yo vote es el que se muestra abajo.

$$\frac{\text{El beneficio neto medio que se deriva para los norteamericanos de la elección de Superior} \times \text{el número de norteamericanos}}{\text{Cien millones}} - \text{los costes para mí y para otros del hecho de que vote}$$

Como hay doscientos millones de norteamericanos, es probable que esta cuenta salga positiva. Será así si la elección de Superior reporta como media a los norteamericanos un beneficio neto que supere en más de la mitad a los costes del hecho de que yo vote. Tendría que ser muy cínico para negar esto. Observaciones similares se aplican a muchos otros bienes públicos, y tanto a los altruistas como a los consecuencialistas. Si un altruista no ignora las probabilidades muy pequeñas, a menudo tendrá una razón moral para hacer una contribución. El beneficio esperado que proporcionaría a los demás sería mayor que los costes de su contribución.

[43] Véase Meehl, Riker y Ordeshook, y Mackie (3).

Puede objetarse que es *irracional* considerar probabilidades muy pequeñas. Cuando nuestros actos no pueden afectar más que a unas pocas personas, puede que sea así. Pero es así porque lo que está en juego es aquí comparativamente bajo. Consideremos los riesgos de causar una muerte accidental. Tal vez sea irracional dedicar algún pensamiento a una probabilidad de una entre un millón de matar a una persona. Pero, si yo fuese ingeniero nuclear, ¿sería irracional al preocuparme por la misma probabilidad de matar a un millón de personas? Esto no es lo que la mayoría de nosotros piensa. Pensamos, correctamente, que tales probabilidades deben ser tenidas en cuenta. Supongamos que los ingenieros nucleares ignoraran todas las probabilidades en el umbral o por debajo del umbral de uno entre un millón. Podría ocurrir entonces que, por cada uno de los muchos componentes de un reactor nuclear, hubiese una probabilidad de uno entre un millón de que cualquier día este componente se estropeara de modo que provocase una catástrofe. Los que diseñan reactores estarían obrando mal, sin lugar a dudas, si ignoraran esas pequeñas probabilidades. Si hubiera muchos reactores, cada uno con muchos de esos componentes, no pasarían demasiados días antes de que el riesgo de uno en un millón hubiese sido pasado un millón de veces. Con seguridad, pronto se produciría una catástrofe.

Cuando lo que está en juego es mucho, no se debería ignorar ninguna probabilidad, por pequeña que fuese. Lo mismo ocurre cuando cada probabilidad va a ser afrontada muchísimas veces. En ambos casos, se debería tomar cada pequeña probabilidad como lo que es, incluyéndose en el cálculo del beneficio esperado. Normalmente podemos ignorar una probabilidad muy pequeña. Pero no deberíamos hacerlo cuando puede afectar a un número enorme de personas, o cuando la probabilidad va a ser afrontada un enorme número de veces. Estos grandes números es como si suprimiesen la pequeñez de la probabilidad.

Algo parecido ocurre si un acto, probablemente o con certeza, va a dar a los demás beneficios muy pequeños. No deberíamos igno-

rar tales beneficios cuando vayan a favor de un enorme número de personas. Este gran número es como si suprimiese la pequeñez de los beneficios. La suma total de los beneficios puede así ser grande.

Estos dos puntos no son igual de convincentes. Los beneficios muy pequeños pueden ser imperceptibles. Y es plausible afirmar que un «beneficio imperceptible» *no* es un beneficio. Pero no es plausible afirmar que una probabilidad muy pequeña *no* es una probabilidad.

28. Ignorar efectos pequeños o imperceptibles

El Tercer Error en matemáticas morales consiste en ignorar las probabilidades muy pequeñas cuando podrían afectar a muchísima gente o podrían ser afrontadas muchísimas veces. Los Errores Cuarto y Quinto son ignorar efectos *muy pequeños* e *imperceptibles* sobre números muy elevados de personas. Son errores similares, y pueden criticarse con los mismos argumentos. Pero los efectos imperceptibles plantean una cuestión suplementaria.

No necesito formular los dos errores. El Cuarto es el mismo que el Quinto, excepto que «muy pequeños» reemplaza a «imperceptibles». Algunos creen que

> (El Quinto Error.) Si determinado acto tiene efectos imperceptibles sobre otras personas, no puede ser moralmente incorrecto *porque* tenga estos efectos. Un acto no puede ser incorrecto a causa de sus efectos sobre otras personas, si ninguna de esas personas pudiera nunca notar ninguna diferencia. De forma similar, si determinado acto tuviera efectos imperceptibles sobre otras personas, estos efectos no podrían hacer de ese acto lo que alguien tuviera que hacer.

Un tipo de efecto imperceptible no se presta a controversia. Yo puedo causarte un perjuicio serio *de un modo* que sea imperceptible. La dosis de radiación que te aplico puede ser la causa desconocida del cáncer que te mate muchos años después. Aunque la causa puede ser desconocida, el efecto es aquí perceptible. Pero en los casos que consideraré, los *efectos* son imperceptibles.

Consideremos en primer lugar una variante de un caso descrito por Glover [44].

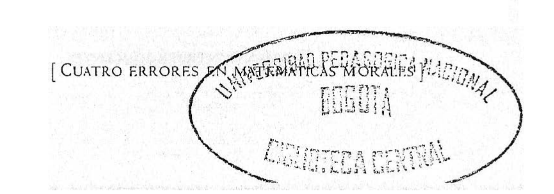
> *Las Gotas de Agua*. Un gran número de hombres heridos yace en el desierto, padeciendo de intensa sed. Nosotros somos un número de altruistas igual de grande, cada uno de los cuales tiene una pinta de agua. Podríamos echar estas pintas dentro de un camión de agua. Lo

[44] Glover (2), pp. 174-5. Glover escribe:
«Puede pensarse que no hay diferencia entre umbrales absolutos y umbrales de discriminación. Algunas personas se sienten tentadas a asimilar el caso de la falta de electricidad al caso de la votación. En el caso de la electricidad, el perjuicio que causo cuando lo extiendo a la comunidad se halla por debajo del umbral de discriminación. Los consecuencialistas que tratan las dos clases de umbral del mismo modo, concluyen que, aparte de los efectos secundarios, no importa si uso o no la electricidad. La sugerencia es que el perjuicio causado cuenta como cero.

Pero contra esto quiero sostener que el perjuicio causado en estos casos debería ser valorado como una fracción de una unidad discriminable, en vez de como cero. Llamemos a esto el *principio de divisibilidad*. Dice que, en casos en que el perjuicio es una cuestión de grado, las acciones por debajo del umbral son incorrectas en la medida en que causan perjuicio, y donde se necesitan cien actos como el mío para causar una diferencia detectable yo he causado el 1/100 de ese perjuicio detectable.

Supongamos que los 100 miembros inermes de una tribu que vive en una aldea, se están comiendo su almuerzo. 100 bandidos armados y hambrientos caen sobre la aldea y cada uno de ellos se apodera a punta de pistola del almuerzo de un miembro de la tribu y se lo come. Luego los bandidos se largan, habiendo causado cada uno de ellos una cantidad de perjuicio discriminable a un miembro de la tribu. A la semana siguiente, los bandidos se ven tentados a hacer otra vez lo mismo, pero se encuentran preocupados por dudas recién descubiertas sobre la moralidad de un asalto como este. Sus dudas son desactivadas por uno de ellos que no cree en el principio de divisibilidad. Entonces asaltan la aldea, atan a los miembros de la tribu y echan un vistazo a su almuerzo. Como esperaban, cada cuenco de comida contiene 100 judías con tomate. El placer que se deriva de una judía con tomate se halla por debajo del umbral de discriminación. Cada bandido, en vez de comerse un plato entero como la semana pasada, coge una judía de cada plato. Se van después de comerse todas las judías, contentos de no haber causado ningún perjuicio, ya que cada uno de ellos no ha causado a cada persona más que un perjuicio que se halla por debajo del umbral. Los que rechazan el principio de divisibilidad tienen que estar de acuerdo.»

Este capítulo, sobre todo mis *Torturadores Inofensivos* de después, deriva enteramente del estímulo de este brillante ejemplo.



llevarían al desierto y repartirían nuestra agua equitativamente entre todos estos hombres heridos. Añadiendo su pinta, cada uno de nosotros permitiría a cada hombre herido beber un poco más de agua —tal vez sólo una gota extra—. Incluso para un hombre muy sediento, cada una de estas gotas extra sería un beneficio muy pequeño. El efecto sobre cada hombre podría ser incluso imperceptible.

Asumamos que el beneficio que se proporciona a cada hombre no sería más que el alivio de una sed intensamente penosa. No habría ningún efecto en la salud de estos hombres. Puesto que los beneficios consistirían simplemente en el alivio del sufrimiento, son la clase de beneficios de la que puede decirse con la máxima plausibilidad que, para ser beneficios en absoluto, tienen que ser perceptibles.

Supongamos en primer lugar que, como los números no son muy grandes, el beneficio que cada uno de nosotros proporcionaría a cada hombre sería, aunque muy pequeño, perceptible. Si incurrimos en el Cuarto Error, pensaremos que tales pequeños beneficios no tienen significado moral. Pensaremos que el que determinado acto fuese a reportar a otros esos pequeños beneficios no podría hacer de este acto lo que alguien debiera hacer. Nos veríamos forzados a concluir que ninguno de nosotros debe añadir su pinta. Esta es sin duda la conclusión incorrecta.

Asumamos a continuación que hay mil hombres heridos y mil altruistas. Si echamos nuestras pintas en el camión del agua, cada uno de nosotros hará que cada hombre herido beba una milésima de pinta extra. Estos hombres podrían notar la diferencia entre *no* beber nada de agua y beber una milésima de pinta. Preguntémonos, por consiguiente, «Si estos hombres bebieran al menos un décimo de pinta, ¿podrían notar el efecto de beber una milésima de pinta extra?». Asumiré que la respuesta es No. (Si la respuesta fuera Sí, necesitaríamos simplemente suponer que hay más altruistas y más hombres heridos. Tiene que haber alguna fracción de pinta cuyo efecto fuese demasiado pequeño para ser perceptible.)

Supongamos que cien altruistas ya han echado su agua en el camión. Cada hombre herido beberá al menos un décimo de pinta. Nosotros somos los otros novecientos altruistas, cada uno de los

cuales podría añadir su pinta. Supongamos a continuación que incurrimos en el Quinto Error. Pensamos que, si determinado acto fuera a tener efectos imperceptibles sobre los demás, estos efectos no podrían hacer de este acto lo que alguien debiera hacer. Si pensamos esto, no podemos explicar por qué cada uno de nosotros debe añadir su pinta.

Puede decirse: «Podemos evitar este problema si redescribimos el efecto de añadir cada pinta. No necesitamos afirmar que esto da a cada uno de los hombres una milésima de pinta. Podríamos afirmar que esto da a un hombre una pinta».

Esta afirmación es falsa. El agua será equitativamente repartida entre todos estos hombres. Cuando yo añado mi pinta, ¿el efecto consiste en que un hombre extra recibe una pinta completa? Si yo no hubiera añadido mi pinta, ¿habría algún hombre que no hubiera recibido nada en vez de recibir una pinta completa? Ninguna de estas dos cosas es cierta. Hay sólo una descripción correcta del efecto de mi acto. Este da a cada uno de los mil hombres una milésima de pinta extra.

Puede decirse a continuación que deberíamos apelar a la Concepción de la Parte-del-Total. Según esta concepción, la parte con la que cada uno contribuye equivale al beneficio que un hombre recibe de una pinta. Pero no podemos apelar a esta concepción puesto que vimos en la Sección 25 que puede implicar tesis absurdas.

A lo que podemos apelar es a una afirmación sobre lo que hacemos *juntos*. Podemos afirmar que

(C10) Cuando (1) la mejor de las consecuencias fuese aquella en la que la gente se beneficiara en mayor medida, y (2) cada uno de los miembros de un determinado grupo pudiese actuar de cierto modo, y (3) ellos beneficiarían a estas otras personas si un número *suficiente* de ellos actuara de ese modo, y (4) ellos beneficiarían a estas personas *en la mayor medida* si *todos* actuaran de ese modo, y (5) cada uno de ellos conoce estos hechos y cree que un número suficiente de ellos actuará de ese modo, entonces (6) cada uno de ellos debe actuar de ese modo.

Cada uno de nosotros podría dar a cada uno de los mil hombres heridos una milésima de pinta extra de agua. Si suficientes de noso-

tros actuamos así, esto beneficiará a cada uno de estos hombres. Y nosotros beneficiaremos a estos hombres en la mayor medida si todos actuamos así. Conocemos estos hechos, y sabemos que suficientes de nosotros —cien— han actuado ya de este modo. (C10) implica correctamente que cada uno de nosotros debe actuar de este modo.

Recordemos ahora el Quinto Error. Según esta concepción, un acto no puede ser correcto o incorrecto, *a causa* de sus efectos sobre otras personas, si estos efectos son imperceptibles. El caso recién descrito refuta esta manera de pensar. Está claro que, en este caso, cada uno de nosotros debería echar su pinta en el camión del agua. Cada uno de nosotros debería propiciar que cada hombre herido bebiese una milésima de pinta extra. Cada uno de nosotros debe afectar a cada hombre herido de esta manera, aunque estos efectos sean imperceptibles. Podemos pensar que, como estos efectos son imperceptibles, cada uno de nosotros no beneficia a nadie. Pero aun si *cada uno* no beneficia a nadie, *juntos* beneficiamos enormemente a estos hombres heridos. Los efectos de *todos* nuestros actos son perceptibles. Aliviamos enormemente la intensa sed de estos hombres.

Los consecuencialistas pueden apelar a varios principios. Pueden por ejemplo pensar que, en algunos casos, la mejor consecuencia de todas no es aquella en la que la gente se beneficia en la mayor medida. Para incluir tales casos, pueden afirmar que

> (C11) Cuando (1) los miembros de un grupo producirían un mejor resultado si un número *suficiente* de ellos actuara de cierto modo, y (2) producirían el resultado *mejor posible* si *todos* ellos actuaran de ese modo, y (3) cada uno de ellos conoce estos hechos y cree que un número suficiente de ellos actuará de ese modo, entonces (4) cada uno de ellos debe actuar de ese modo.

Los no-consecuencialistas piensan que, en ciertos casos, deberíamos tratar de producir el mejor resultado posible. En estos casos, pueden apelar a (C11). Como antes, en algunos casos, (C11) no nos da por sí misma la respuesta correcta. Tendríamos

que añadir una afirmación más complicada. Ignoraré estas complicaciones [46].

Como hice ver en la Sección 26, hay dos clases de casos en que necesitamos apelar a los efectos, no sólo de actos singulares, sino de conjuntos de actos. Necesitamos hacer esta apelación cuando (1) los efectos de nuestros actos están sobredeterminados, o (2) nos enfrentamos con problemas de coordinación. Estamos considerando ahora casos en que (3) el acto de cada persona tendrá efectos imperceptibles en otras personas. Esto puede ser una tercera clase de casos en que necesitamos apelar a los efectos de conjuntos de actos. Que *necesitemos* hacer esta apelación depende en parte de la respuesta a otra cuestión.

29. ¿Puede haber perjuicios y beneficios imperceptibles?

Puede objetarse: «Afirmas que cada uno de los mil altruistas debería echar su pinta, puesto que así se beneficiarían en la mayor medida los hombres heridos. Esta afirmación es falsa. Supongamos que uno de los altruistas no echa su pinta. ¿Se beneficiarán menos los hombres heridos? No. Beberán una cantidad de agua ligeramente menor. Pero este efecto es imperceptible. Como el efecto es imperceptible, el beneficio para estos hombres no puede ser menor».

Esta objeción da por sentado que no puede haber beneficios imperceptibles. Si hacemos esta asunción, nos enfrentamos a parte de un problema más amplio, al que se denomina de varias maneras: el *Problema Sorites*, la *Paradoja de Wang*, o la *Paradoja del Montón*.

En nuestro caso, el beneficio consiste en el alivio de una sed intensamente penosa. Si cada hombre recibe una pinta de agua, su sed se hará menos penosa. Su sufrimiento será menos malo. Nuestro problema es el siguiente. Damos por sentado que

> (A) El sufrimiento de una persona no puede hacerse *imperceptiblemente* mejor o peor. El sufrimiento de una persona no puede

[46] Véase una vez más Regan.

hacerse ni menos malo ni peor, si esta persona no pudiera de ningún modo notar diferencia alguna.

Y es plausible asumir que

(B) *Al menos tan malo como y no peor que* son, cuando los aplicamos a sufrimientos y dolores, relaciones *transitivas*. De modo que, si el dolor de una persona en el Resultado (2) es al menos tan malo como lo era en el Resultado (1), y su dolor en el Resultado (3) es al menos tan malo como lo era en el resultado (2), su dolor en el Resultado (3) tiene que ser al menos tan malo como lo era en el Resultado (1).

Cien altruistas han echado ya sus respectivas pintas. Cada uno de los hombres heridos beberá al menos un décimo de pinta. No notarían el efecto de una milésima de pinta extra. En diferentes resultados posibles diferentes números de altruistas echarán después sus pintas al camión. Refirámonos a estos resultados citando el número del que contribuye. Así, si nadie más contribuye, esto producirá el Resultado 100.

Supongamos que contribuye un altruista más. Cada herido beberá más agua, pero la cantidad será tan pequeña que no podrá notarlo. Según (A), la sed de cada hombre no puede hacerse menos penosa. El sufrimiento de cada hombre en el resultado 101 tiene que ser al menos tan malo como lo habría sido en el resultado 100. Supongamos a continuación que un segundo altruista añade su pinta. Como antes, ninguno de los hombres puede notar esta diferencia. Según (A), la sed de cada hombre no puede hacerse menos penosa. La sed de cada hombre en el Resultado 102 tiene que ser al menos tan mala como lo habría sido en el Resultado 101. Según (B), el sufrimiento de cada hombre en el Resultado 102 tiene que ser al menos tan malo como lo habría sido en el *Resultado 100*. Las mismas afirmaciones son de aplicación si un tercer altruista contribuye. En el Resultado 103, el sufrimiento de cada hombre tiene que ser al menos tan malo como lo habría sido en el Resultado 100. Estas afirmaciones se aplican a cada altruista extra que contribuye. Supongamos que todos contribuimos. Estaríamos ante el Resultado 1000, en el que cada hombre bebe una pinta completa. (A) y (B) juntos implican que el sufrimiento de cada hombre tiene que ser al menos tan malo como lo habría sido en el Resultado 100. Beber una pinta entera, en vez de tan sólo un décimo, no puede hacer nada para aliviar el sufrimiento que le produce la sed a cada hombre. Como esta conclusión es absurda, tenemos que rechazar (A) o (B).

¿Qué hacer? Yo rechazo (A). Pienso que el sufrimiento de una persona puede hacerse menos penoso, o menos malo, en una cantidad demasiado pequeña para ser notada. El sufrimiento de una persona es peor, en el sentido que tiene relevancia moral, si a esta persona su sufrimiento le importa más, o tiene un deseo más fuerte de que el sufrimiento cese. Pienso que a alguien le puede importar su sufrimiento ligeramente menos, o puede tener un deseo más débil de que su sufrimiento cese, aunque no pueda notar ninguna diferencia. Más en general, puede haber perjuicios imperceptibles, y beneficios imperceptibles. En muchos otros tipos de casos se ha demostrado que las personas incurren en errores muy pequeños cuando informan de la naturaleza de sus experiencias. ¿Por qué deberíamos dar por sentado que no pueden cometer tales errores cuando se trata de la intensidad de su deseo de que cese algún dolor?

Supongamos que rechazas estas afirmaciones, y continuas aceptando (A). Entonces tienes que rechazar (B). Para evitar la absurda conclusión alcanzada antes, tienes que admitir que, cuando se aplican a dolores y sufrimientos, *al menos tan malo como* y *no peor que* no son relaciones transitivas. Y rechazar (B) tiene implicaciones como las de rechazar (A). Ahora tienes que admitir que tus actos pueden ser incorrectos, a causa de sus efectos en el dolor de alguna otra persona, aunque *ninguno* de tus actos haga peor el dolor de esta persona. Pueden tener este efecto *juntos*. Cada acto puede ser incorrecto, aunque sus efectos sean imperceptibles, porque sea uno de un conjunto de actos que, juntos, hace el dolor de esta persona mucho peor.

Consideremos

Los Viejos Malos Días. Mil torturadores tienen en su poder a mil víctimas. Al comienzo de cada día, cada una de las víctimas ya siente un ligero dolor. Cada uno de los torturadores gira el interruptor de

cierto instrumento unas mil veces. Cada giro del interruptor afecta al dolor de alguna víctima de un modo imperceptible. Pero, después de que cada torturador ha girado su interruptor mil veces, el resultado es que ha infligido a su víctima un dolor severo.

Supongamos que cometes el Quinto Error. Piensas que un acto no puede ser incorrecto a causa de sus efectos en otras personas, si estos efectos son imperceptibles. Entonces tienes que concluir que, en este caso, ningún giro que se le dé al interruptor es malo. Ninguno de estos torturadores obra nunca mal. Esta conclusión es absurda.

¿Por qué obran mal los torturadores? Una explicación apela al efecto total de lo que cada torturador hace. Cada uno de ellos gira el interruptor mil veces. Estos actos, tomados en bloque, infligen un severo dolor a su víctima.

Consideremos a continuación

Los Torturadores Inofensivos. En los Viejos Malos Días, cada torturador infligía un severo dolor a una víctima. Ahora han cambiado las cosas. Cada uno de los mil torturadores aprieta un botón, y con ello gira un interruptor una vez en cada uno de los mil instrumentos. Las víctimas sufren el mismo severo dolor. Pero ninguno de los torturadores hace el dolor de cada víctima perceptiblemente peor.

¿Podemos apelar aquí al efecto total de lo que cada torturador hace? Esto depende en parte de si rechazamos (A), creyendo que el dolor de alguien puede hacerse imperceptiblemente peor. Si creyéramos esto, podríamos afirmar: «Al apretar el botón, cada torturador hace que cada víctima sufra ligeramente más. El efecto en cada una es ligero. Pero, como cada torturador incrementa el sufrimiento de mil víctimas, impone sobre ellas una gran cantidad total de sufrimiento. Como las víctimas sufren justo como sufrían en los Viejos Malos Días, estos torturadores están obrando justo tan mal como solían hacerlo. En los Viejos Malos Días, cada torturador imponía una gran cantidad de sufrimiento a una víctima. Cada uno de los Torturadores Inofensivos impone a estas mil víctimas una cantidad total de sufrimiento igualmente grande».

Supongamos en cambio que aceptamos (A), creyendo que los dolores no pueden hacerse imperceptiblemente peores. Entonces tenemos que admitir que cada uno de los Torturadores Inofensivos no hace sufrir más a nadie. Según nuestra concepción, ninguno de los torturadores perjudica a nadie.

Aun si ninguno de ellos perjudica a nadie, los torturadores sin duda están obrando mal. Si no podemos apelar a los efectos de lo que cada torturador hace, tenemos que apelar a lo que los torturadores hacen juntos. Aunque ninguno de ellos causa ningún dolor, todos juntos imponen gran sufrimiento a mil víctimas. Podemos afirmar

(C12) Cuando (1) el resultado sería peor si la gente sufriese más, y (2) cada uno de los miembros de determinado grupo podría actuar de cierto modo, y (3) harían sufrir a otras personas si *un número suficiente* de ellos actuara de ese modo, y (4) harían sufrir a esas personas *en la mayor medida* si *todos* actuaran de ese modo, y (5) cada uno de ellos no sólo conoce estos hechos sino que además piensa que un número suficiente de ellos actuará de ese modo, entonces (6) cada uno estaría obrando mal si actuara de ese modo.

Alguien puede objetar de nuevo: «En el caso de los Torturadores Inofensivos, (4) no es verdadera. Estos torturadores no harán sufrir a sus víctimas *en la mayor medida* si *todos* ellos giran el interruptor una vez. Supongamos que uno de ellos no girara ningún interruptor. Ninguna de las víctimas notaría ninguna diferencia. Como un dolor no puede hacerse menos malo imperceptiblemente, las víctimas *no* sufrirían menos si uno de los torturadores no actuara».

Como señalé, esta objeción origina el bien conocido Problema Sorites [47]. Si aceptamos (A), nuestra respuesta a esta objeción tiene que involucrar una solución a este problema. Pero como este problema no sólo es difícil de resolver, sino que además genera cuestiones que no tienen nada que ver con la ética, no lo discutiré aquí [48].

[47] Véase Hare (3).

[48] (Nota añadida en 1987.) En su artículo en *Ethics*, julio de 1986, B. Gruzalski demuestra que mi discusión original de este problema era confusa. Véase también Dummett, Peacocke (1), y Forbes (2).

Si aceptamos (A), nuestra objeción a los Torturadores Inofensivos tiene que ser complicada y además resolver el Problema Sorites. Si rechazamos (A), nuestra objeción podría ser simple. Podríamos afirmar que cada uno de los torturadores inflige a las víctimas una gran cantidad total de sufrimiento.

De estas dos explicaciones, ¿cuál es mejor? Aunque rechazáramos (a), podríamos estar equivocados dando la explicación más simple. Que esto sea así depende de la respuesta a otra pregunta. Consideremos

> *El Torturador Individual*. Una mañana, un solo torturador aparece en el trabajo. Por casualidad sucede que, por causas naturales, cada víctima está ya sufriendo verdaderamente un dolor severo. Este dolor es más o menos tan malo como sería después de que los interruptores hubieran sido girados quinientas veces. Conociendo este hecho, el Torturador Individual aprieta el botón que hace girar el interruptor una vez en todas las máquinas. El efecto es el mismo que en los días en que todos los torturadores se ponían manos a la obra. Más precisamente, el efecto es justo como el que se produce cuando se gira cada interruptor la vez quinientas y una. El Torturador Individual sabe que este es el efecto. Sabe que no está haciendo que el dolor de ninguna víctima sea perceptiblemente peor. Y sabe que no forma parte de ningún grupo que esté haciendo esto conjuntamente.

¿Está obrando mal el Torturador Individual? Supongamos que pensáramos que no. Entonces no podemos apelar a la objeción más simple en el caso en que actúan todos los torturadores. No podemos afirmar que cada uno está obrando mal porque está imponiendo a otros una gran cantidad total de sufrimiento. Si es por eso por lo que cada uno obra mal, el Torturador Individual tiene que estar obrando mal. Él actúa del mismo modo, y con los mismos efectos. Si pensamos que el Torturador Individual no está obrando mal, tenemos que plantear la otra objeción en el caso en que actúan todos los demás torturadores. Tenemos que afirmar que cada uno está actuando mal porque es un miembro de un grupo que conjuntamente inflige un gran sufrimiento a sus víctimas.

Yo creo que el Torturador Individual *está* obrando mal. ¿Qué diferencia moral se derivaría de que él produjera malos efectos en colaboración con otros agentes o con la Naturaleza? [49]. Por eso prefiero, en ambos casos, apelar a los efectos de actos singulares.

Algunos no están de acuerdo. Aunque pensemos que puede haber perjuicios y beneficios imperceptibles, tal vez sea mejor, por ello, apelar a lo que los grupos hacen juntos. Esta apelación es menos controvertida.

(Si el Torturador Individual *no* está obrando mal, puede ser injusto afirmar que algunos de nosotros cometemos *cinco* errores en matemáticas morales. Según esta concepción, el Quinto Error es simplemente un caso especial del Segundo Error. Pero esto rara vez importa a efectos prácticos.)

En esta sección he inquirido si puede haber perjuicios y beneficios imperceptibles. Me inclino a contestar Sí. Si la respuesta es No, tenemos que abandonar la afirmación de que, cuando se aplican a perjuicios y beneficios, *al menos tan malo como* y *no peor que* son relaciones transitivas. He mostrado también que importa poco qué respuesta aceptemos. En cualquiera de los dos casos, tenemos que abandonar lo que llamo el Quinto Error. Tenemos que abandonar la concepción de que un acto no puede ser correcto o incorrecto *a causa* de sus efectos en otras personas, si estos efectos son imperceptibles.

30. Sobredeterminación

Volvamos ahora a las pintas de agua y los hombres heridos. Añadamos algunos detalles a este caso. Supongamos que, antes de que el camión del agua sea llevado hasta esos hombres, llegas tú, con otra pinta. Los hombres heridos necesitan más que una única pinta. Tras beber esta pinta su sed intensamente penosa no sería aliviada

[49] Supongamos que el agua que les cae a los heridos no procediese de otros agentes sino de la lluvia. ¿Suprimiría esto la razón que tengo para añadir mi pinta?

del todo. Pero el camión del agua sólo puede contener mil pintas. Ahora está lleno. Si añades tu pinta, esto lo único que significará es que una pinta se eche a perder derramándose.

Tú no tienes ninguna razón moral para añadir tu pinta, puesto que esto lo único que causará es que se desperdicie una pinta. Según (C10), deberías añadir tu pinta si esto te convirtiera en miembro de un grupo que, conjuntamente, beneficia a otras personas. Podemos pensar que, si añades tu pinta, *no* eres miembro del grupo que conjuntamente beneficia a los heridos. Estos heridos pueden beber algo de tu pinta. Y tú actúas del mismo modo que los otros altruistas. Pero podríamos afirmar, «A diferencia de los otros altruistas, tú no das a cada herido una milésima extra de una pinta de agua. Tu acto no tiene repercusión en la cantidad de agua que estos hombres reciben».

Las cosas no son tan simples. Si tú añades tu pinta, esto será un caso de los que conllevan sobredeterminación. Es cierto que si no hubieras contribuido no habría habido diferencia alguna en lo que respecta a la cantidad de agua que beben los hombres. Pero, como has contribuido, pasa lo mismo con cada uno de los otros altruistas. Es cierto que si uno de esos altruistas no hubiese contribuido esto no habría supuesto ninguna diferencia en la cantidad de agua que los hombres beben. El camión del agua no habría estado lleno cuando tú llegaste, y tu pinta lo habría llenado. Lo que es cierto de ti es cierto de cada uno de los otros altruistas. Es por tanto cierto que tú *eres* un miembro del grupo que, conjuntamente, beneficia a los heridos.

Una vez más tenemos que apelar a lo que los agentes saben, o tienen razón para creer. Supongamos que los otros altruistas no tenían razón alguna para pensar que tú llegarías, con tu pinta extra. Cada uno debería haber echado su pinta. Esto es así porque cada uno tenía una buena razón para creer que sería un miembro de un grupo del que es verdadero no sólo (1) que conjuntamente iba a beneficiar a los heridos, sino también (2) que beneficia a esos hombres *en la mayor medida* si *todos* sus componentes echan sus pintas. Cuando tú llegas, sabes que el camión del agua está lleno. No tienes ninguna razón para contribuir, ya que sabes que *no* serías un

miembro de tal grupo. Si tú contribuyes, serás en cambio miembro de un grupo *que es demasiado grande*. Deberíamos afirmar

> (C13) Supongamos que hay cierto grupo que, al actuar de cierto modo, beneficiará conjuntamente a otras personas. Si alguien cree que este grupo es, o sería si él se uniera, *demasiado grande*, no tiene ninguna razón moral para unirse a ese grupo. Un grupo es *demasiado grande* si es cierto que, si uno o más de sus miembros no hubieran actuado, esto no habría reducido el beneficio que este grupo proporciona a otras personas.

Si añades tu pinta, esto hará de este grupo de altruistas un grupo demasiado grande. Si *no* añades tu pinta, este grupo *no* será demasiado grande. Este es un caso límite especial. (C13) también incluye los casos más comunes en que determinado grupo es ya demasiado grande.

31. ALTRUISMO RACIONAL

El Quinto Error en matemáticas morales es la creencia de que los efectos imperceptibles no pueden ser moralmente significantes. Este es un error muy serio. Cuando todos los Torturadores Inofensivos actúan, cada uno de ellos está obrando *muy* mal. Esto es cierto aun cuando ninguno haga a nadie sentirse peor. Lo mismo podría ser cierto de nosotros. Deberíamos dejar de pensar que un acto no puede ser incorrecto *a causa* de sus efectos sobre otras personas, si este acto no hace a nadie sentirse peor. Cada acto nuestro puede ser *muy* malo a causa de sus efectos en otras personas, aunque ninguna de esas personas pueda jamás notar ninguno de esos efectos. Nuestros actos pueden *en conjunto* hacer que esas personas se sientan muchísimo peor.

El Cuarto Error es igualmente serio. Si creemos que los efectos insignificantes pueden ser moralmente ignorados, a menudo podemos hacer que la gente se sienta mucho peor. Recordemos el Dilema del Pescador. Donde hay exceso de capturas, o reservas en declive, puede ser mejor para cada uno tratar de capturar más, pero peor para cada uno si todos lo hacen. Consideremos

Cómo los Pescadores Provocan un Desastre. Hay muchos pescadores, que se ganan la vida pescando cada uno por su cuenta en un gran lago. Si ningún pescador restringe sus capturas, en las próximas temporadas, pocas, capturará más pescado. Pero con ello reducirá la captura total en un número mucho más grande. Como hay muchos pescadores, si ninguno restringe sus capturas afectará sólo de modo insignificante a la cantidad capturada por cada uno de los demás. Los pescadores creen que tales efectos insignificantes pueden ser moralmente ignorados. Puesto que esto es lo que creen, aunque nunca hagan lo que creen incorrecto, no restringen sus capturas. Con ello cada uno aumenta sus propias capturas, pero provoca una reducción mucho más grande en las capturas totales. Puesto que todos ellos obran así, el resultado es un desastre. Después de unas cuantas temporadas, todos capturan poquísimo pescado. No pueden alimentar a sus hijos, ni alimentarse a sí mismos.

Si estos pescadores conocieran los hechos, tuvieran suficiente altruismo, y evitaran el Cuarto Error, se librarían de este desastre. Cada uno sabe que, si no restringe sus capturas, esto de algún modo será mejor para él, hagan lo que hagan los demás. Y cada uno sabe que, si obra así, los efectos en cada uno de los otros serán insignificantes. Pero los pescadores no deberían creer que estos efectos insignificantes pueden ser moralmente ignorados. Deberían creer que obrar así es incorrecto.

Como antes, hay dos modos en que podríamos explicar por qué estos actos son incorrectos. Podríamos apelar al efecto total del acto de cada persona. Cada pescador sabe que, si no restringe sus capturas, capturará más pescado, pero reducirá la captura total en un número mucho mayor. Para conseguir una pequeña ganancia para sí mismo, impone a los demás una pérdida total mucho más grande. Podríamos afirmar que tales actos son incorrectos. Esta afirmación no asume que puede haber perjuicios y beneficios imperceptibles. Es por tanto menos controvertida que la correspondiente afirmación sobre lo que hace cada uno de los Torturadores Inofensivos.

Nuestra alternativa es apelar a lo que estos pescadores hacen juntos. Cada pescador sabe que, si él y todos los demás no restringen sus capturas, juntos se impondrán a sí mismos una gran pérdida total. Los altruistas racionales pensarían que estos actos son incorrectos. Evitarían este desastre.

Puede decirse: «Y también los egoístas racionales. Cada uno sabe que, si no restringe sus capturas, entra a formar parte de un grupo que se impone a sí mismo una gran pérdida. Es irracional obrar así incluso en términos del propio interés». Como defenderé en el próximo capítulo, esta afirmación no está justificada. Cada uno sabe que, si no restringe sus capturas, esto será *mejor* para él. Esto es así, hagan lo que hagan los demás. Cuando alguien hace lo que sabe será mejor para él mismo, no puede afirmarse que su acto sea irracional en términos del propio interés.

Recordemos a continuación

El Dilema del Trabajador Viajero. Supongamos que vivimos en los suburbios de una gran ciudad. Podemos ir y volver al trabajo en coche o en autobús. Como no hay carriles-bus, el tráfico extra llega a colapsar a los autobuses tanto como a los coches. Por tanto, podríamos saber que lo siguiente es cierto: cuando la mayoría de nosotros va en coche, si alguno de nosotros va en coche en vez de ir en autobús se ahorrará con ello algún tiempo, pero impondrá a los demás una pérdida de tiempo total mucho mayor. Este efecto se dispersaría. Cada uno podría provocar que cien de los otros se retrasaran veinte segundos, o que mil de los otros se retrasaran dos segundos. La mayoría de nosotros consideraría tales efectos como tan insignificantes que podrían ser moralmente ignorados. Pasaríamos a creer entonces que, en este Dilema del Trabajador Viajero, incluso un altruista racional podría elegir, de manera justificada, ir en coche en vez de ir en autobús. Pero si la mayoría de nosotros hiciéramos esta elección todos nosotros nos retrasaríamos mucho tiempo cada día.

Los altruistas racionales evitarían este resultado. Como antes, podrían apelar o a los efectos de lo que cada persona hace o a los efectos de lo que todos juntos hacemos. Cada uno se ahorra algún tiempo a costa de imponer a los demás una pérdida total de tiempo mucho mayor. Podríamos afirmar que es incorrecto obrar así, aunque los efectos en cada uno de los demás sean insignificantes.

Podríamos en cambio afirmar que este acto es incorrecto porque los que actúan así, juntos, imponen a cada uno de los demás una gran pérdida de tiempo. Si aceptamos cualquiera de estas afirmaciones, y tenemos suficiente altruismo, resolveríamos el Dilema del Trabajador Viajero, ahorrándonos mucho tiempo cada día.

Razonamientos parecidos se aplican a otros casos innumerables. Por poner otro ejemplo, consideremos los dispositivos que purifican los gases que emiten nuestros coches. Pensaríamos que es incorrecto ahorrarnos el coste de reparar este dispositivo si, como consecuencia, impusiéramos una gran polución atmosférica a alguna otra persona individual. Pero muchos de nosotros no pensarían que esto es incorrecto si simplemente incrementara de manera insignificante o imperceptible la polución atmosférica sufrida por cada una de muchísimas personas. Este sería el efecto real en muchas grandes ciudades. Podría ser mucho mejor para todos nosotros si ninguno de nosotros causara tal polución. Pero, para creer que actuamos mal, muchos de nosotros necesitamos cambiar nuestra manera de pensar. Tenemos que dejar de creer que un acto no puede ser incorrecto, a causa de sus efectos en otras personas, si estos efectos son insignificantes o imperceptibles.

Cuando cambian las condiciones, puede que necesitemos hacer algunos cambios en el modo en que pensamos acerca de la moralidad. He estado defendiendo un cambio así. La Moralidad del Sentido Común funciona del mejor modo en comunidades pequeñas. Cuando hay pocas personas, si damos o imponemos a otros beneficios o perjuicios totales grandes, tenemos que estar afectando a otras personas de maneras significativas, con lo que habría razones para la gratitud o el resentimiento. En comunidades pequeñas es verosímil afirmar que no podemos haber perjudicado a otros si no hay nadie con una queja obvia o con razones para sentirse molesto por lo que hemos hecho.

Hasta este siglo, la mayor parte de la humanidad vivía en pequeñas comunidades. Lo que cada uno hacía podía afectar sólo a unos cuantos. Pero ahora han cambiado las condiciones. Ahora cada uno de nosotros puede, de muchísimas maneras, afectar a un sinnúmero

de otras personas. Podemos tener efectos reales aunque pequeños en miles o millones de personas. Cuando estos efectos se dispersan ampliamente, pueden llegar a ser insignificantes, o imperceptibles. Ahora representa una gran diferencia si seguimos creyendo que no podemos haber perjudicado o beneficiado mucho a otros a no ser que haya personas con razones obvias para el resentimiento o la gratitud. Mientras sigamos pensando así, aunque nos preocupemos de los efectos sobre los demás, puede que fracasemos a la hora de resolver muchos Dilemas del Prisionero bastante serios. Para conseguir pequeños beneficios para nosotros mismos, o para nuestras familias, cada uno de nosotros puede negar a los demás beneficios totales mucho más grandes, o imponerles perjuicios totales mucho más grandes. Puede que pensemos que esto es permisible porque los efectos sobre cada uno de los demás van a ser insignificantes o imperceptibles. Si es esto lo que pensamos, lo que hagamos será con frecuencia mucho peor para todos nosotros.

Si nos preocupáramos lo suficiente de los efectos en los otros, y cambiáramos nuestra concepción moral, resolveríamos tales problemas. No basta con preguntar, «¿Perjudicará mi acto a otras personas?». Aunque la respuesta sea No, puede que mi acto sea todavía incorrecto, a causa de sus efectos. Los efectos que tendrá cuando sea considerado en sí mismo puede que no sean sus solos efectos relevantes. Debería preguntar, «¿Será mi acto uno de un conjunto de actos que, *juntos*, perjudicarán a otras personas? La respuesta puede ser Sí. Y el daño a otros puede ser muy grande. Si esto es así, puede que yo esté obrando *muy* mal, como los Torturadores Inofensivos. Tenemos que aceptar esta conclusión si nuestro interés por los demás va a proporcionarnos soluciones a la mayoría de los muchos Dilemas del Prisionero que afrontamos: la mayoría de los muchos casos en que, si cada uno de nosotros en vez de ninguno hace lo que será mejor para sí mismo —o para su familia, o para los que quiere— esto será peor, y con frecuencia *mucho* peor, para todos.

4

TEORÍAS QUE SON DIRECTAMENTE CONTRAPRODUCENTES

A menudo nos enfrentamos a Dilemas del Prisionero de Muchas Personas. Es con frecuencia cierto que, si cada uno en vez de ninguno de nosotros hace lo que será mejor para sí mismo, o su familia, o aquellos a los que quiere, esto será peor para todos nosotros. Si cada uno de nosotros está dispuesto a actuar de este modo, estos casos plantean un problema práctico. A no ser que cambie algo, el resultado será peor para todos nosotros.

Este problema tiene dos clases de solución: política y psicológica. De las soluciones psicológicas las más importantes son las soluciones morales. Como defendí, hay muchos casos en que necesitamos una solución moral.

Describí cuatro de estas soluciones. Estas nos son proporcionadas por cuatro motivos: formalidad, renuencia a «gorronear», querer satisfacer el Test Kantiano, y altruismo suficiente. Hay dos formas de cada solución moral. Cuando uno de estos motivos lleva a alguien a hacer la elección altruista, lo que esta persona hace puede ser o no ser peor para ella. Esta distinción plantea profundos interrogantes. Expondré, simplemente, lo que dan por sentado mis argumentos. Según todas las teorías verosímiles del propio interés, lo que va a favor de nuestros intereses depende en parte de cuáles

son nuestros motivos o deseos. Si tenemos motivos morales, puede que por ello no sea cierto que la elección altruista vaya a ser peor para nosotros. Pero esto podría ser cierto. Incluso si lo fuera, todavía podríamos hacer esta elección.

Estoy rechazando aquí cuatro afirmaciones. Algunos dicen que nadie hace lo que cree que será peor para él. Esto ha sido refutado a menudo. Otros dicen que lo que cada uno hace es, por definición, lo mejor de todo para él. Para decirlo como el economista, esto «maximizará su utilidad». Puesto que es simplemente una definición, no puede ser falso. Pero aquí es irrelevante. Simplemente no tiene que ver con lo que va a favor del propio interés a largo plazo de una persona. Otros dicen que la virtud es siempre recompensada. A no ser que haya una vida después de la muerte, esto ha sido refutado también. Otros dicen que la virtud es su propia recompensa. Según la Teoría de la Lista Objetiva, ser moral y actuar moralmente pueden ser de las cosas que hagan que nuestras vidas vayan mejor. Pero, según las versiones plausibles de esta teoría, podría haber casos en que actuar moralmente sería, vistas las cosas en su conjunto, peor para alguien. Actuar moralmente podría privar a la persona de demasiadas de las otras cosas que hacen que nuestras vidas vayan mejor.

Para volver a mis propias afirmaciones: muchos Dilemas del Prisionero necesitan soluciones morales. Para llegar a estas soluciones, tenemos que estar directamente dispuestos a hacer la elección altruista. Hay dos formas de cada solución moral. Una forma suprime el dilema. En estos casos, puesto que tenemos algún motivo moral, no es cierto que vaya a ser peor para cada uno el que haga la elección altruista. Pero en otros casos esto es aún cierto. Incluso en tales casos, podríamos hacer esta elección. Cada uno podría hacer, por razones morales, lo que sabe que va a ser peor para él.

A menudo necesitamos soluciones morales de esta segunda forma. Llamémoslas *abnegadas*. Resuelven el problema práctico. El resultado es mejor para todos. Pero no suprimen el dilema. Queda un problema teórico.

El problema es este: podemos tener razones morales para hacer la elección altruista. Pero será mejor para cada uno si hace la elec-

ción que le beneficia a sí mismo. La moralidad entra en conflicto con el propio interés. Cuando los dos entran en conflicto, ¿qué es racional hacer?

Según la teoría del Propio Interés, es la elección que le beneficia a uno mismo la que es racional. Si creemos en PI, seremos ambivalentes respecto de las soluciones morales abnegadas. Pensaremos que, para lograr tales soluciones, todos tenemos que actuar irracionalmente.

Muchos autores se resisten a esta conclusión. Algunos afirman que las razones morales no son más débiles que las razones del propio interés. Otros, con más atrevimiento, afirman que son más fuertes. Según su modo de ver, es la elección que le beneficia a uno mismo la que es irracional.

Puede que este debate parezca irresoluble. ¿Cómo podemos pesar estas dos clases de razones la una contra la otra? Las razones morales son, desde luego, supremas en el sentido moral. Pero las razones del propio interés son supremas en el sentido del propio interés. ¿Dónde podemos encontrar una escala neutral?

32. En los Dilemas del Prisionero, ¿falla PI en sus propios términos?

Se ha afirmado que no necesitamos una escala neutral. Hay un sentido en que, en los Dilemas del Prisionero, la teoría del Propio Interés es contraproducente. Se ha afirmado que, puesto que es así, las razones morales son superiores a las razones del propio interés, aun en términos del propio interés.

Como hemos visto, PI podría ser individual e indirectamente contraproducente. Podría ser peor para alguien que nunca fuera abnegado. Pero esto no es así en los Dilemas del Prisionero. Los malos efectos son aquí producidos por actos, no por disposiciones. Y está claro qué elección será mejor para cada persona. Es verdadero de cada uno que, si hace la elección altruista, esto será con certeza peor para él. PI le dice a cada uno que haga la elección que le beneficia a sí mismo. Y, hagan lo que hagan los otros, será mejor

para cada uno si él mismo hace esta elección. PI no es aquí individualmente contraproducente. Pero, en el sentido definido en la Sección 22, PI es directa y colectivamente contraproducente. Si todos seguimos PI con éxito, esto será para cada uno peor que si nadie lo hace.

¿Muestra esto que, si todos nosotros seguimos PI, somos irracionales? Podemos comenzar con una pregunta más fácil. Si creemos en PI, ¿fracasaría nuestra teoría incluso en sus propios términos?

Podríamos contestar: «No. La búsqueda particular del propio interés es mejor para cada uno. ¿Por qué es PI colectivamente contraproducente? Sólo porque la búsqueda del propio interés es peor para otros. Esto no la convierte en un fracaso. No es benevolencia».

Si nos guiamos por el propio interés, deberemos desde luego deplorar los Dilemas del Prisionero. Estos no son los casos que les encantan a los economistas clásicos, casos en que cada uno gana si todos buscan el propio interés. Podríamos decir: «En esos casos, PI no sólo funciona sino que también aprueba la situación. En los Dilemas del Prisionero, PI aún funciona. Todavía gana cada uno por su búsqueda particular del propio interés. Pero como cada uno pierde aún más por los actos de los otros guiados por el propio interés, PI, aquí, condena la situación».

Esto puede parecer una evasión. Cuando sea peor para cada persona que todos nosotros persigamos el propio interés, puede parecer que la teoría del Propio Interés *debería* condenarse a sí misma. Supongamos que en otro grupo, enfrentándose a los mismos dilemas, todos hacen la elección altruista. Podrían decirnos: «Pensáis que somos irracionales. Pero salimos más beneficiados que vosotros. Lo hacemos mejor hasta en los términos del propio interés».

Nosotros podríamos responder: «Eso es sólo un juego de palabras. Vosotros "lo hacéis mejor" sólo en el sentido de que salís más beneficiados. Cada uno de vosotros lo está *haciendo* peor en los términos del propio interés. Cada uno está haciendo lo que sabe que será peor para él». Y aun podríamos añadir: «Lo que es peor para cada uno de nosotros es que, en nuestro grupo, no haya tontos. Cada uno de vosotros tiene mejor suerte. Aunque tu irracionalidad es mala para ti, ganas aún más con la irracionalidad de los otros».

Ellos podrían responder: «En parte tenéis razón. Cada uno de nosotros lo *hace* peor en los términos del propio interés. Pero, aunque *cada uno* lo haga peor, *nosotros* lo hacemos mejor. Esto no es un juego de palabras. Cada uno de nosotros sale mejor parado a causa de lo que nosotros *hacemos*».

Esta sugerencia es más prometedora. Volvamos al Caso, más simple, de las Dos Personas. Cada persona podría o bien beneficiarse a sí misma (E) o bien proporcionar a la otra algún beneficio mayor (A). Los resultados serían los que se muestran abajo.

		Tú: haces E	Tú: haces A
Yo	hago E	Lo tercero mejor para cada uno	Lo mejor para mí, lo peor para ti
	hago A	Lo peor para mí, lo mejor para ti	Lo segundo mejor para los dos

Para asegurarnos de que la elección de uno no puede afectar a la del otro —lo que podría producir reciprocidad— supongamos que no nos podemos comunicar. Si yo hago A en vez de E, eso será entonces peor para mí. Lo cual es así hagas tú lo que hagas. Y lo mismo ocurre contigo. Si los dos hacemos A en vez de E, cada uno está por tanto haciéndolo peor en los términos del propio interés. La sugerencia es que *nosotros* lo estamos haciendo mejor.

Lo que hace a esto prometedor es que establece un contraste entre *cada uno* y *nosotros*. Como hemos visto, lo que es falso de cada uno puede ser verdadero de nosotros. Puede ser verdadero, por ejemplo, que, aunque *cada uno* por su cuenta no perjudique a nadie, *nosotros juntos* perjudiquemos a otras personas. Si los dos hacemos A y no E, ¿es verdad que, aunque cada uno lo hace peor en términos del propio interés, los dos juntos lo estamos haciendo mejor?

Podemos usar esta prueba. La teoría del Propio Interés proporciona a cada persona un determinado fin. Cada persona lo hace mejor, en los términos de PI, si, de aquellos actos que son posi-

bles para ella, hace lo que ocasiona que su fin PI-dado sea mejor logrado. *Nosotros* lo hacemos mejor en términos de PI si, de aquellos actos que son posibles para nosotros, hacemos lo que ocasiona que los fines *de cada uno* PI-dados sean mejor logrados. Esta prueba parece justa. Podría demostrar que, si cada uno lo hace lo mejor que puede en términos de PI, nosotros juntos no lo podríamos hacer mejor.

Cuando medimos el éxito sólo cuentan los fines *últimos*. Supongamos que estamos intentando rascarnos la espalda. El fin último de cada uno podría ser que deje de picarle. Entonces haríamos lo mejor si cada uno rascara la espalda del otro. Pero podríamos ser contorsionistas: el fin último de cada uno podría ser rascarse la espalda *a sí mismo*. Si nos rascáramos la espalda el uno al otro, lo estaríamos haciendo entonces peor.

¿Qué fin último proporciona a cada persona la teoría del Propio Interés? ¿Que se promuevan sus intereses, o que *ella misma* promueva sus intereses? Según la teoría del Propio Interés, si los intereses de alguien son promovidos por sí mismo, esta persona está actuando racionalmente. Por eso puedo reformular mi pregunta. ¿Cuál es el fin último que PI le adjudica a cada persona? ¿Que sean promovidos sus intereses o que actúe racionalmente?

En la Sección 3 defendí la siguiente respuesta. Como todas las teorías de la racionalidad, PI le adjudica a cada persona el fin formal de que actúe racionalmente. Pero, según PI, este fin formal no es, como tal, un fin sustantivo. PI adjudica a cada persona un fin sustantivo último: que su vida vaya, para ella, lo mejor posible. Según la Teoría Hedonista del propio interés, ser racional y obrar racionalmente no forman parte de este fin. Ambos son simples medios. Según otras teorías del propio interés, ser racional y obrar racionalmente no son meros medios. Los dos son, independientemente de sus efectos, partes del fin sustantivo último que PI adjudica a cada persona. Pero esto no sería cierto cuando fuesen, en su conjunto, peores para alguien.

Podemos imaginar una teoría que da a cada persona este fin sustantivo: que sus intereses sean promovidos *por sí misma*. Alguien que crea en esta teoría podría malinterpretar burdamente a Nietzsche, y

valorar «la más feroz autosuficiencia» [53]. Si los dos hiciéramos A en vez de E, estaríamos haciéndolo peor en estos términos paranietzscheanos. Los intereses de cada cual se promoverían mejor. Pero ninguno de los dos promovería sus intereses por sí mismo, de manera que el fin paranietzscheano sería peor logrado.

Si los dos hiciésemos A en lugar de E, ¿estaríamos haciéndolo mejor en términos de PI? Estaríamos haciendo que los intereses particulares de cada uno se promovieran mejor. En este aspecto, estaríamos haciéndolo mejor en términos de PI, haciendo que el fin PI-dado de cada uno sea mejor logrado. Según la Teoría Hedonista del propio interés, esto responde completamente a mi pregunta. Según esta teoría, PI afirma que cada acto es un mero medio. El fin es siempre el efecto en la vida consciente de uno. (Las «bestias rubias» de Nietzsche fueron, se dice, leones. Pero, también para los leones, actuar es un medio. Prefieren comer lo que otros matan.)

Según algunas otras teorías del propio interés, tiene que decirse más. De acuerdo con PI, si los dos hacemos A en lugar de E, los dos estamos actuando irracionalmente. Cada uno está haciendo lo que sabe será peor para él. Según algunas teorías del propio interés, ser racional y obrar racionalmente son parte del fin que PI adjudica a cada uno. Según estas teorías, algunos aparentes Dilemas del Prisionero no son verdaderos Dilemas. Discuto estos casos en la nota [54].

[53] Véase Nagel (1), p. 127.

[54] Según estas teorías del propio interés, ser racional y actual racionalmente son, sean sus efectos los que puedan ser, mejores para cada persona. Pero son sólo dos de las cosas que son buenas para nosotros. Por tanto, podrían ser, tomadas las cosas en conjunto, peores para nosotros. Esto sería así cuando, si fuéramos racionales y actuáramos racionalmente, nos hiciésemos perder a nosotros mismos demasiadas de las otras cosas que son buenas para nosotros. Cuando esto ocurre así, ser racional y obrar racionalmente no son parte del fin último que nos es dado por PI. Pero esto no ocurre así en los Dilemas del Prisionero. En estos casos si cada persona hace lo que PI dice que es racional, esto será mejor para ella.

Supongamos que aceptamos una de estas teorías. ¿Cómo deberíamos contestar a mi pregunta? Si los dos hacemos A más bien que PI, ¿lo estamos haciendo mejor en los términos de PI? ¿Estamos haciendo que los fines PI-dados de cada uno sean mejor logrados? La respuesta es: En un sentido, Sí; en otro, No. Hacemos que cada una de

En los Dilemas genuinos, si ambos hacemos A en vez de E, lo estamos haciendo mejor en términos de PI. Estamos causando que el fin PI-dado a cada uno sea mejor logrado. Esto es así según todas las teorías del propio interés. Lo hacemos *mejor* en términos de PI si hacemos lo que PI nos dice que *no* hagamos.

¿Muestra esto que PI fracasa en sus propios términos? Puede parecerlo. Y es tentador contrastar PI con la moralidad. Podríamos decir, «La teoría del Propio Interés alimenta el conflicto, al decirle a cada persona que trabaje contra las otras. Así es como, si todos buscan el propio interés, esto puede ser malo para todos. Donde la teoría del Propio Interés divide, la moralidad une. Nos dice que trabajemos juntos —que hagamos lo mejor que *nosotros* podamos—. Incluso según la escala que nos proporciona el propio interés, la moralidad, por tanto, gana. Esto es lo que aprendemos de los Dilemas del Prisionero. Si dejáramos de guiarnos por el propio interés y nos volviéramos morales, lo haríamos mejor incluso en términos del propio interés» [55].

Este argumento fracasa. *Nosotros* lo hacemos mejor, pero *cada uno* lo hace peor. Si los dos hacemos A en vez de E, *nosotros* producimos el resultado mejor para cada uno, pero *cada uno* produce el resultado peor para sí mismo. Haga el otro lo que haga, sería mejor para cada uno hacer E. En los Dilemas del Prisionero, el problema es este: ¿debería *cada persona* hacerlo lo mejor que pueda para sí misma? ¿O deberíamos *nosotros* hacerlo lo mejor que podamos para cada persona? Si *cada persona* hace lo que es mejor para sí misma, *nosotros* lo hacemos peor de lo que podríamos para cada persona. Pero *nosotros* lo hacemos mejor para cada persona sólo si *cada persona* lo hace peor de lo que podría para sí misma.

Lo cual es sólo un caso especial de un problema más amplio. Consideremos cualquier teoría sobre lo que tenemos razón para hacer. Podrían darse casos en que, si cada uno lo hace mejor en términos de esta teoría, nosotros lo hacemos peor, y viceversa. Llamemos a tales casos *Dilemas Cada Uno-Nosotros*. Una teoría puede producir tales Dilemas aunque no se preocupe de qué es lo que favorece nuestros intereses.

Las teorías consecuencialistas no pueden producir esos Dilemas. Como vimos en la Sección 21, esto ocurre porque estas teorías son *neutrales respecto del agente*, adjudicando a todos los agentes fines comunes.

Si una teoría produce Dilemas Cada Uno-Nosotros, puede que no sea obvio lo que esto demuestra. Reconsideremos la teoría del Propio Interés. Esta teoría le dice a cada persona que haga lo mejor que pueda para sí misma. Estamos discutiendo casos en que, si

nuestras vidas vaya, en un sentido, mejor. Hacemos el resultado mejor para cada uno. Pero hacemos que cada una de nuestras vidas vaya peor, en otro sentido. Estamos obrando irracionalmente. Según estas teorías sobre el propio interés, es en sí mismo malo para nosotros obrar irracionalmente, cualesquiera que puedan ser los efectos.

Ahora tenemos que preguntar si, en conjunto, hacemos que cada una de nuestras vidas vaya mejor. La respuesta depende de los detalles del caso. Si ambos hacemos A más bien que PI, ¿hasta qué punto hacemos el resultado mejor para cada uno? Si lo hacemos muchísimo mejor, esto compensará el mal rasgo de que estamos obrando irracionalmente. Si producimos un resultado sólo levemente mejor, esto no compensará este mal rasgo. Recordemos el ejemplo original. Si los dos guardamos silencio, cada uno librará al otro de diez años de prisión, al coste de sumar dos años al tiempo que uno mismo pasará en prisión. El efecto neto es que cada uno se librará de ocho años en prisión. Este efecto es mejor para cada uno. Lo logramos al coste de obrar irracionalmente. Si es malo para nosotros obrar irracionalmente de esta manera, ¿es esto tan malo como pasar ocho años en prisión? Podríamos decir, «Es peor». Según este modo de ver las cosas, el caso no es un verdadero Dilema. No es verdadero que si cada uno más bien que ninguno hace lo que va a ser mejor para él mismo, eso será peor para ambos. Según nuestro modo de ver las cosas, será mejor para cada uno confesar antes que permanecer callado, desde el momento en que así se librará de dos años en prisión. Pero no sería mejor para cada uno si los dos confesaran en vez de permanecer en silencio. Si ambos permanecen callados, cada uno de nosotros nos ahorramos ocho años de cárcel, al coste de obrar irracionalmente. Según la concepción mencionada ahora mismo, esto será peor para cada uno. Obrar racionalmente de esta manera es peor para cada uno que pasarse ocho años en prisión.

En este ejemplo, esta respuesta puede parecer absurda. Pero supongamos que, al coste de obrar irracionalmente, le ahorramos a cada uno ocho horas de cárcel, o sólo ocho minutos. Entonces sería menos absurdo afirmar que el caso no es un verdadero Dilema.

[55] Muchos autores han argumentado en estos términos. Véase por ejemplo Baier (1) y (3), y Gauthier (3) y (4).

todos buscamos el propio interés, hacemos lo que es peor para cada uno. La teoría del Propio Interés es aquí directa y colectivamente contraproducente. Pero no podemos asumir que esto sea un defecto. ¿Por qué razón debería ser PI colectivamente exitosa? ¿Por qué no es suficiente con que, a nivel individual, PI tenga éxito?

Podríamos decir: «Una teoría no puede aplicarse sólo a un único individuo. Si es racional para mí hacer lo que va a ser lo mejor para mí, tiene que ser racional para todos hacer lo que va a ser lo mejor para cada uno. Cualquier teoría aceptable tiene que ser, por consiguiente, exitosa a nivel colectivo».

Esto lleva consigo una confusión. Llamemos a una teoría *universal* cuando se aplica a todos, *colectiva* cuando afirma tener éxito a nivel colectivo. Algunas teorías tienen ambos rasgos. Un ejemplo es la moral kantiana. Esta le dice a cada persona que haga sólo lo que podría racionalmente querer que hiciera todo el mundo. Los planes o los programas de conducta de cada uno tienen que probarse a nivel colectivo. Para un kantiano, la esencia de la moralidad es el movimiento de *cada uno* a *nosotros*.

A nivel colectivo —como una respuesta a la pregunta, «¿Cómo deberíamos actuar todos?— la teoría del Propio Interés *se condenaría* a sí misma. Supongamos que estamos eligiendo qué código de conducta será fomentado públicamente, y enseñado en las escuelas. PI nos diría aquí que votásemos contra ella. Si estamos eligiendo un código colectivo, la elección guiada por el propio interés sería alguna versión de la moralidad.

PI es universal, se aplica a todos. Pero PI no es un código colectivo. Es una teoría de la racionalidad individual. Esto contesta la pregunta más fácil que hice antes. En los Dilemas del Prisionero, PI tiene éxito individualmente. Puesto que es sólo colectivamente contraproducente, PI no falla en sus propios términos. PI no se condena a sí misma.

34. Dilemas Intertemporales

Muchas malas teorías no se condenan a sí mismas. Así que la pregunta más amplia continúa abierta. En tales casos, ¿qué es racional hacer?

Puede ayudar que presentemos otra teoría común. Es la que le dice a cada uno que haga lo que mejor vaya a lograr sus fines presentes. Llamémosla la teoría del *fin Presente*, o *P*. Supongamos que, en algún Dilema del Prisionero, mi fin es el resultado mejor para mí. De acuerdo con P, es entonces la elección que me beneficia a mí la que es racional. Si mi fin es beneficiar a otros, o pasar el Test Kantiano, será la elección altruista la racional. Si mi fin es hacer lo que los otros hacen —tal vez porque no quiero ser un «gorrón»— no es seguro qué elección es la racional. Esto dependerá de mis creencias acerca de lo que los otros hacen.

Como estas observaciones enseñan, P puede entrar en conflicto con PI. Aquello que logrará mis fines presentes lo mejor posible puede ir en contra de mi propio interés a largo plazo. Ya que las dos teorías pueden entrar en conflicto, los que defienden PI tienen que rechazar P.

Ellos podrían señalar que, incluso a nivel individual, P puede ser directamente contraproducente. Puede producir *Dilemas Intertemporales*. Éstos se les plantearán más comúnmente a los que se preocupan menos por su futuro. Supongamos que soy una persona así, y que, en diferentes períodos de tiempo, tengo diferentes fines. En cada período de tiempo yo podría o bien (1) hacer lo que conseguirá mis fines presentes lo mejor posible, o bien (2) hacer lo que conseguirá lo mejor posible, o me pondrá en condiciones de conseguir lo mejor posible, todos mis fines a lo largo del tiempo. Según P, yo debería siempre hacer (1) en vez de (2). Sólo así yo, *en cada período de tiempo*, lo haré lo mejor que pueda en términos de P. Pero *a lo largo del tiempo* puede que yo entonces lo haga peor, en estos mismos términos. A lo largo del tiempo, puede que yo tenga menos éxito en lograr mis fines en cada período de tiempo.

Aquí va un ejemplo trivial pero, en mi caso, verdadero. En cada período de tiempo yo lograré mejor mis fines presentes si no gasto energía siendo ordenado. Pero si no soy nunca ordenado esto puede hacer que yo en cada período de tiempo posterior consiga menos. Y mi falta de orden puede frustrar lo que yo trataba de conseguir la primera vez. Entonces será cierto, como tristemente lo es, que no ser nunca ordenado hace que yo en cada período de tiempo consiga menos.

Los que creen en la teoría del Propio Interés pueden apelar a tales casos. Podrían decir: «La teoría del fin Presente es aquí directamente contraproducente. Incluso en los términos de P, PI es superior. La elección del propio interés es (2). Si haces siempre (2) en vez de (1), lograrás con más efectividad tus fines en cada período de tiempo. Si sigues PI, haces lo mejor incluso en términos de P».

Como la similar defensa de la moralidad, este argumento falla. Si yo sigo PI, lo hago mejor *a lo largo del tiempo*. Pero *en cada período de tiempo* lo hago peor. Si siempre hago (2), estoy en cada período de tiempo haciendo aquello que logrará *menos* efectivamente los fines que yo entonces tengo. (1) es lo que los logrará mejor.

Esta distinción puede ser difícil de entender. Supongamos que yo siempre hago (1) en lugar de (2). Será entonces un hecho que, *a lo largo del tiempo*, lograré con *menos* efectividad los fines que yo tengo *en cada período de tiempo*. Si esto es un hecho, ¿cómo puede ser también verdad que, en cada período de tiempo, yo lograré con *más* efectividad mis fines *en ese período de tiempo*? Para ver cómo es esto posible, podemos recordar el Dilema Interpersonal. Donde dijimos antes «nosotros» ponemos ahora «yo a lo largo del tiempo», y donde dijimos «cada uno» ponemos ahora «yo en cada período de tiempo». En el Dilema Interpersonal, nosotros lo hacemos mejor *para cada uno* si cada uno lo hace peor de lo que podría *para sí mismo*. En el Dilema Intertemporal, yo lo hago mejor a lo largo del tiempo *en cada período de tiempo* sólo si en cada período de tiempo lo hago peor de lo que yo *entonces* podría.

Como sugieren estas afirmaciones, los Dilemas Cada Uno-Nosotros son un caso especial de un problema aún más amplio. Llamémosles *Dilemas de la Relatividad de la Razón*. PI produce Dilemas Cada Uno-Nosotros porque sus razones son *relativas al agente*. Según PI, yo puedo tener una razón para hacer justo lo que tú puede que tengas una razón para deshacer. P produce Dilemas Intertemporales porque sus razones son *relativas al tiempo*. Según P, yo puedo tener ahora una razón para hacer lo que más tarde tendré una razón para deshacer.

P puede ser intertemporalmente contraproducente. Pero P no afirma que tenga éxito a nivel *intertemporal*. Es una teoría sobre aquello para hacer lo cual en cada período de tiempo tenemos razones. Incluso en los Dilemas Intertemporales, P tiene éxito en cada período de tiempo. Si yo siempre sigo P, haciendo (1) en vez de (2), estoy haciendo en cada período de tiempo lo que logrará mejor mis fines en ese período de tiempo. Como P es una teoría de lo que tenemos razones para hacer en cada período de tiempo, no falla aquí en sus propios términos. P no se condena a sí misma.

Un teórico del Propio Interés tiene que afirmar que, sin embargo, debemos rechazar P. Podría decir: «Cualquier teoría aceptable tiene que ser intertemporalmente exitosa. No sirve de defensa que P no reivindique tal éxito. Esto meramente revela que P tiene un defecto estructural. Si una teoría es intertemporalmente contraproducente, esto basta para demostrar que debería ser rechazada».

Puede que estas afirmaciones no hagan nada para apoyar a PI. Si P se refuta por el hecho de que es intertemporalmente contraproducente, ¿por qué no queda refutada PI dado el hecho de que es interpersonal —o colectivamente— contraproducente? Y si es una buena respuesta que PI no afirma ser colectivamente exitosa, ¿por qué no puede dar una respuesta similar el teórico del fin Presente?

Como indican estas observaciones, la teoría del Propio Interés puede ser desafiada desde dos direcciones. Esto la hace más difícil de defender. Las respuestas a uno de los desafíos pueden socavar las respuestas al otro.

Un desafío viene de las teorías morales. El otro, de la teoría del fin Presente. Hay varias versiones de esta teoría. La versión más simple es la *Teoría Instrumental*. Según ella, aquello que cada persona tiene más razones para hacer es lo que sea que consiga mejor el cumplimiento de sus fines presentes. Esta teoría toma los fines del agente como dados, y discute sólo de medios. Ningún fin se declara irracional. Cualquier fin puede proporcionar buenas razones para actuar.

Otra versión de P es la *Teoría Deliberativa*. Esta apela, no a los fines presentes y reales del agente, sino a los fines que tendría

ahora si conociera los hechos relevantes y pensara con claridad. Según esta teoría, si un fin no sobreviviera al proceso de deliberación, entonces no nos proporcionaría una buena razón para actuar.

Una tercera versión de P critica los fines de un segundo modo. Según esta teoría, aunque sobrevivan al proceso de deliberación, ciertas clases de fin son intrínsecamente irracionales, y no pueden proporcionar buenas razones para actuar. Lo que cada persona tiene más razones para hacer es lo que sea que logre mejor esos fines presentes suyos que no son irracionales. Esta es la *teoría Crítica del fin Presente*.

En todas sus versiones, P a menudo entra en conflicto con la teoría del Propio Interés, Alguien puede conocer los hechos y estar pensando con claridad, pero tener fines que sabe que van contra su propio interés a largo plazo. Y podemos pensar que algunos de esos fines no son irracionales. Algunos ejemplos podrían ser: beneficiar a los demás, descubrir verdades y crear belleza. Podemos concluir que la búsqueda de estos fines no es menos racional que la búsqueda del propio interés. Según esta concepción, buscar tales fines no es irracional aunque el agente sepa que está actuando en contra de su propio interés a largo plazo.

Un teórico del Propio Interés tiene que rechazar estas afirmaciones. Tiene que insistir en que las razones para actuar no pueden ser relativas al tiempo. Podría decir: «La fuerza de una razón se prolonga en el tiempo. Como yo *tendré* una razón para promover mis fines futuros, tengo una razón para hacerlo así *ahora*». Esta afirmación está en el corazón de la teoría del Propio Interés.

Muchos teóricos morales hacen una segunda afirmación. Piensan que ciertas razones no son relativas al agente. Podrían decir: «La fuerza de una razón puede prolongarse, no sólo en el tiempo, sino también en diferentes vidas. Así, si *yo* tengo una razón para aliviar mi dolor, esta es una razón también para ti. *Tú* tienes una razón para aliviar *mi* dolor».

El teórico del Propio Interés hace la primera afirmación, pero rechaza la segunda. Puede encontrar difícil de defender las dos mitades de esta posición. En respuesta al moralista, puede preguntar, «¿Por qué debería *yo* asignar un peso a fines que no son *míos*?». Pero un teórico del fin Presente puede preguntar, «¿Por qué debería yo asignar un peso *ahora* a fines que no son míos *ahora*?». El teórico del Propio Interés puede contestar con una apelación a los Dilemas Intertemporales, aquellos en que la teoría del fin Presente es intertemporalmente contraproducente. Pero entonces puede a su vez ser desafiado con los Dilemas Interpersonales, aquellos en que su propia teoría es colectivamente contraproducente. Un moralista podría decir: «El argumento de la teoría del Propio Interés nos lleva más allá de esta teoría. Correctamente entendido, es un argumento a favor de la moralidad».

En la Segunda Parte seguiré esta línea de pensamiento. Pero primero se debería discutir algo más. A nivel interpersonal, el contraste *no* se da entre la teoría del Propio Interés y la moralidad.

36. Cómo la Moralidad del Sentido Común es directamente contraproducente

Como di a entender en la Sección 22, la teoría del Propio Interés no es la única teoría que puede producir Dilemas Cada Uno-Nosotros. Tales casos pueden ocurrir cuando (a) determinada teoría T es relativa al agente, dando a agentes diferentes fines diferentes, (b) el logro de los fines T-dados de cada persona depende parcialmente de lo que los otros hagan, y (c) lo que cada uno haga no afectará a lo que estos otros hagan. Estas condiciones con frecuencia se dan para la Moralidad del Sentido Común.

La mayor parte de nosotros considera que hay ciertas personas hacia las que tenemos obligaciones especiales. Se trata de las personas con las que nos hallamos en ciertas relaciones —tales como nuestros hijos, nuestros padres, amigos, benefactores, alumnos, pacientes, clientes, colegas, miembros de nuestro propio Sindicato, aquellos a los que representamos o nuestros conciudadanos. Consideramos que debemos intentar preservar a todas estas personas de ciertas clases de perjuicios, y que debemos intentar darles

ciertas clases de beneficios—. La Moralidad del Sentido Común consiste en gran medida en tales obligaciones.

Cumplir estas obligaciones tiene prioridad sobre ayudar a los desconocidos. Esta prioridad no es absoluta. No debo salvar a mi hijo pequeño de un corte o de un cardenal antes que salvarle la vida a un desconocido. Pero debo salvar a mi hijo pequeño de algún perjuicio antes que salvar a un desconocido de un perjuicio *algo* mayor. Mi deber para con mi hijo pequeño no queda anulado con tal de que pudiese hacer un bien algo mayor en otro lugar.

Cuando trato de proteger a mi hijo pequeño, ¿cuál debería ser mi fin? ¿Debería ser simplemente que no resultara perjudicado? ¿O debería ser más bien que *yo* le librara del perjuicio? Si tú tuvieras una probabilidad mejor de salvarle del perjuicio, me equivocaría si insistiera en que el intento lo tengo que hacer yo. Esto enseña que mi fin debería adoptar la forma más simple.

Podemos enfrentarnos a *Dilemas de Padres*. Consideremos

Caso Uno. Nosotros no nos podemos comunicar. Pero cada uno podría o bien (1) librar a su propio hijo pequeño de algún perjuicio, o bien (2) librar al hijo pequeño de otro de algún perjuicio algo más grande. Los resultados se muestran abajo.

		Tú: haces (1)	Tú: haces (2)
Yo	hago (1)	Nuestros dos hijos pequeños sufren el perjuicio mayor	El mío no sufre ningún perjuicio, el tuyo sufre los dos
Yo	hago (2)	El mío sufre los dos perjuicios, el tuyo no sufre ninguno	Los dos sufren el perjuicio menor

Como no podemos comunicarnos, ninguna elección que tome uno afectará a la que tome el otro. Si la finalidad de cada uno debiera ser que su niño no resultara perjudicado, cada uno debería hacer aquí (1) en vez de hacer (2). Porque de esta manera cada uno de nosotros se aseguraría de que su hijo resulta perjudicado en menor medida. Esto es así haga lo que haga el otro. Pero si los dos

hacen (1) en lugar de (2) nuestros dos hijos pequeños serán perjudicados más.

Consideremos a renglón seguido los beneficios que yo debiera tratar de dar a mi hijo pequeño. ¿Cuál debería ser aquí mi finalidad? ¿Debería insistir en que fuese *yo* el que beneficiara a mi hijo, aunque esto fuera peor para él? Algunos responderían siempre No. Pero esta respuesta puede ser demasiado radical. Trata el cuidado parental como un mero medio. Podemos pensarlo mejor. Podemos estar de acuerdo en que, con algunas clases de beneficio, mi fin debería adoptar la forma más simple. Debería consistir simplemente en que el resultado fuese mejor para mi hijo. Pero puede haber otros tipos de beneficio que mi hijo debería recibir *de mí*.

Con ambos tipos de beneficio, podemos hacer frente a los Dilemas de Padres. Consideremos

Caso Dos. Nosotros no nos podemos comunicar. Pero cada uno podría o bien (1) beneficiar a su propio hijo pequeño, o bien (2) beneficiar algo más al hijo del otro. Los resultados se muestran abajo.

		Tú: haces (1)	Tú: haces (2)
Yo	hago (1)	El tercero mejor para nuestros dos hijos	El mejor para el mío, el peor para el tuyo
Yo	hago (2)	El peor para el mío el mejor para el tuyo	El segundo mejor para los dos

Si mi finalidad debiera ser aquí que el resultado fuese mejor para mi hijo, yo debería de nuevo hacer (1) en lugar de (2). Y lo mismo vale para ti. Pero si ambos hacemos (1) en lugar de (2) esto será peor para nuestros hijos. Comparemos

Caso Tres. Nosotros no nos podemos comunicar. Pero yo podría o bien (1) permitirme dar a mi hijo pequeño algún beneficio, o bien (2) permitirte beneficiar al tuyo algo más. Tú tienes

las mismas alternativas con respecto a mí. Los resultados se muestran debajo.

		Tú: haces (1)	Tú: haces (2)
Yo	hago (1)	Cada uno puede dar a su hijo algún beneficio	Yo puedo beneficiar al mío al máximo, tú puedes beneficiar al tuyo al mínimo
	hago (2)	Yo puedo beneficiar al mío al mínimo, tú puedes beneficiar al tuyo al máximo	Cada uno puede beneficiar a su hijo más

Si mi finalidad debiera ser aquí que yo beneficiase a mi hijo, yo debería una vez más hacer (1) antes que (2). Y lo mismo vale para ti. Pero si ambos hacemos (1) en lugar de (2) cada uno podrá beneficiar menos a su hijo. Nótese la diferencia entre estos dos ejemplos. En el Caso Dos estamos interesados en lo que *ocurre*. La finalidad de cada uno de nosotros es que el resultado sea mejor para su hijo. Esta es una finalidad que el otro puede hacer que se logre de manera directa. En el Caso Tres estamos interesados en lo que nosotros *hacemos*. Como mi finalidad es que *yo* beneficie a mi hijo pequeño, tú no puedes, en mi nombre, hacerlo. Pero sí que me puedes ayudar a hacerlo. De este modo puedes ayudarme indirectamente a que mi fin se logre.

No es probable que ocurran Dilemas de Padres de Dos Personas. Pero con frecuencia hacemos frente a Versiones de Muchas Personas. A menudo es cierto que, si todos en vez de ninguno damos prioridad a nuestros propios hijos, esto o bien será peor para todos ellos o bien permitirá a cada uno beneficiar menos a sus propios hijos. Así que hay muchos bienes públicos: resultados que beneficiarían a nuestros hijos tanto si ayudamos a producirlos como si no. Puede ser verdadero de cada padre el que, si no ayuda, esto será mejor para sus propios hijos. Lo que ahorre —en dinero, tiempo o energía— lo puede emplear en beneficiar sólo a sus propios

hijos. Pero si ningún padre ayudase a producir este bien público esto sería para todos nuestros hijos peor que si todos ayudásemos. En otro caso corriente, tal como el Dilema del Pescador, cada uno podría o bien (1) aumentar sus propias ganancias, o bien (2) (por autocontrol) aumentar en mayor medida las ganancias de los demás. Aquí será verdadero de cada uno que, si hace (1) en lugar de (2), podrá beneficiar más a sus hijos. Y esto es así, hagan lo que hagan los demás. Pero si todos hacen (1) en vez de (2) cada cual podrá beneficiar menos a sus hijos. Estas son sólo dos de las maneras en que estos casos pueden ocurrir. Hay muchas otras.

Observaciones similares se aplican a todas las obligaciones parecidas —como las que tenemos con los alumnos, los pacientes, los clientes o los electores—. Con todas las obligaciones como éstas, hay incontables Versiones de Muchas Personas como mis tres Dilemas de Padres. Son tan corrientes y tan variadas como los Dilemas del Prisionero de Muchas Personas. Como acabamos de ver, con frecuencia tendrán la misma causa.

Aquí tenemos otro modo en que esto podría ser cierto: supongamos que, en el Dilema del Prisionero original, son nuestros abogados lo que tienen que elegir. Esto plantea el *Dilema del Abogado del Prisionero*. Si los dos abogados dan prioridad a sus propios clientes esto será para los dos clientes peor que si ninguno lo hace. Cualquier Dilema del propio interés puede de este modo dar pie a un dilema moral. Si un grupo se enfrenta al primero, otro puede en consecuencia enfrentarse al segundo. Esto puede ser así si cada miembro del segundo grupo debe dar prioridad a algunos miembros del primero. Una afirmación similar es de aplicación cuando grupos diferentes, tales como naciones, se enfrentan a un Dilema del propio interés. La mayoría de los gobiernos considera que debe dar prioridad a los intereses de sus ciudadanos. Hay varios modos en que, si todos los gobiernos antes que ninguno dieran prioridad a sus propios ciudadanos, esto sería peor para todos sus ciudadanos. El problema viene de *la asignación de prioridad*. No cambia nada si se la asigna a uno mismo o a otros.

Todos mis ejemplos llevan consigo perjuicios o beneficios. Pero el problema puede surgir por otras partes de la Moralidad del

Sentido Común. Puede surgir siempre que esta moralidad dé a diferentes personas diferentes deberes. Supongamos que cada uno pudiera o bien (1) cumplir algunos de sus propios deberes, o bien (2) permitir a otros cumplir más de los suyos. Si todos antes que ninguno dieran prioridad a sus propios deberes, cada uno podría cumplir menos. Los que se ocupan de la Deontología pueden enfrentarse a Dilemas Cada Uno-Nosotros.

37. Las cinco partes de una teoría moral

¿Qué es lo que muestran tales casos? La mayoría de nosotros acepta alguna versión de la teoría que denomino Moralidad del Sentido Común. Según ella, hay determinadas cosas que cada uno de nosotros debe tratar de conseguir. Estas son lo que denomino nuestros *fines morales*. Seguimos con éxito esta teoría moral cuando cada uno hace lo que, de los actos que son posibles para él, mejor consigue sus fines morales. En mis casos es cierto que, si todos antes que ninguno seguimos con éxito esta teoría, con ello ocasionaremos que los fines morales de cada uno sean peor conseguidos. Nuestra teoría moral es aquí directa y colectivamente contraproducente. ¿Constituye esto una objeción?

Vamos a empezar con una pregunta de menor alcance. ¿Podríamos revisar nuestra teoría, para que no fuese contraproducente? Si no cabe tal revisión, la nuestra puede ser la mejor teoría posible. En primer lugar debemos identificar la parte de nuestra teoría que es contraproducente.

Una parte de una teoría moral puede incluir *actos satisfactorios*, con la asunción de *conformidad plena*. Llamemos a esta parte *Teoría del Acto Ideal*. Esta dice lo que todos nosotros deberíamos tratar de hacer, simplemente con la asunción de que todos nosotros lo intentamos, y todos tenemos éxito. Llamemos a esto *lo que todos nosotros deberíamos hacer idealmente*. «Todos» aquí no significa «todo el que vive». Significa «los miembros de algún grupo».

Como defendí en el capítulo 1, no basta con decidir lo que todos nosotros debemos hacer idealmente. Tenemos que tomar en consideración estos cuatro hechos:

(a) a menudo no tenemos certeza respecto a cuáles serán los efectos de nuestros actos;
(b) algunos de nosotros podemos obrar mal;
(c) nuestros actos no son los únicos efectos de nuestros motivos;
(d) cuando sentimos remordimiento, o nos echamos la culpa unos a otros, esto puede afectar a lo que después hagamos, y tener otros efectos.

Nuestra teoría moral puede, por consiguiente, tener las cinco partes que se muestran abajo.

	Todos	Cada Uno
Actos satisfactorios	Nuestra *Teoría del Acto Ideal*, que dice lo que todos deberíamos hacer idealmente, cuando sabemos que todos tendremos éxito	Nuestra *Teoría del Acto Práctico*, que dice lo que cada uno de nosotros debe hacer, dados (a) y (b)
Motivos	Nuestra *Teoría del Motivo Ideal*, que dice qué motivos deberíamos tener todos idealmente, dados (a) y (c)	Nuestra *Teoría del Motivo Práctico*, que dice qué motivos debería tener cada uno de nosotros, dados (a), (b) y (c)
Culpa y remordimiento		Nuestra *Teoría de la Reacción*, que dice cuáles son los actos por los que cada uno debe ser culpado y debería sentir remordimiento, dados (a), (b), (c) y (d)

Cuando estamos decidiendo qué creer, deberíamos considerar en primer lugar nuestra Teoría del Acto Ideal. Al preguntar lo que todos deberíamos hacer idealmente, estamos preguntando cuáles deberían ser nuestros fines morales últimos. Estos conforman el fundamento de nuestra teoría moral. Las otras partes de nuestra teoría son lo que necesitamos afirmar, supuestos nuestros fines morales últimos, cuando consideramos los cuatro hechos establecidos arriba.

38. Cómo podemos revisar la Moralidad del Sentido Común para que no sea contraproducente

Supongamos que aceptamos alguna versión de la Moralidad del Sentido Común. En mis ejemplos, lo que es verdadero es esto. Si *todos* nosotros seguimos *con éxito* nuestra teoría moral, será directamente contraproducente. Lo que es contraproducente es nuestra Teoría del Acto Ideal. Si debiéramos revisar nuestra teoría, esta sería la primera parte que deberíamos revisar.

Llamemos a nuestra teoría *M*, y a su versión revisada *R*. Una de las afirmaciones de R es

> (R1) Cuando M es contraproducente, todos deberíamos idealmente hacer lo que vaya a ocasionar que los fines M-dados *de cada uno* sean mejor logrados.

Así, en todos mis Dilemas de Padres todos deberíamos idealmente hacer (2) antes que (1). Esto produciría un mejor resultado para todos nuestros hijos, y permitiría a cada cual beneficiar a sus propios hijos en mayor medida.

(R1) revisa nuestra Teoría del Acto Ideal. Si revisamos esta parte de nuestra teoría, seremos conducidos de forma natural a revisar el resto.

Consideremos en primer lugar nuestra Teoría del Acto Práctico. Esta describe lo que cada uno de nosotros debe hacer, dados los hechos (a) de que a menudo no sabemos cuáles van a ser los efectos de nuestros actos, y (b) de que algunos de nosotros obraremos mal.

Volvamos al caso de un bien público que beneficiaría a nuestros hijos. Un bien tal es la conservación de recursos escasos. Supongamos que somos los pescadores pobres del Dilema del Pescador, tratando de capturar lo suficiente para alimentar a nuestros hijos. Como hay captura abusiva de pescado, las reservas se están agotando. Es verdad que si cada persona no restringe su captura, esto será ligeramente mejor para sus propios hijos. Estarán ligeramente mejor alimentados. Esto es así hagan lo que hagan los demás. Pero si ninguno de nosotros restringe sus capturas esto será mucho peor para todos nuestros hijos que si todos nosotros restringimos nuestras capturas. Todos nuestros hijos estarán mucho peor alimentados. Según (R1), todos deberíamos idealmente restringir nuestras capturas. Si alguno deja de hacerlo, (R1) deja de aplicarse. Pero sería natural hacer esta afirmación suplementaria: cada uno debería restringir sus capturas a condición de que un número *suficiente* de los demás lo haga también.

¿Qué cuenta como suficiente? Hay una respuesta natural a esta pregunta. Consideremos cualquier bien público que pudiera beneficiar a nuestros hijos, y que vaya a ser otorgado sólo en el caso de que haya contribuciones voluntarias. Asumamos, por simplicidad, que no hay umbral superior por encima del cual las contribuciones se desperdiciarían. Nuestros hijos se beneficiarían en la mayor medida posible si todos nosotros contribuimos. Supongamos que cada uno de nosotros sabe que algunos padres no contribuirán. Tiene que haber un número mínimo *k* tal que, si k o más padres contribuyen, esto será mejor para los hijos de cada contribuyente que si ninguno contribuyera. Si sólo contribuye *uno*, esto será *peor* para sus hijos que si no contribuyera. Si *todos* contribuimos, esto será *mejor* para todos nuestros hijos que si nadie contribuyera. En algún sitio entre *uno* y *todos* tiene que estar el número k donde tenga lugar el cambio de *peor* a *mejor* [57].

[57] Véase Schelling (2).

El número k tiene dos características especiales: (1) Si *k o más* contribuyen, cada contribuyente esta componiendo una trama cuyo efecto neto es beneficiar a sus propios hijos. Los hijos de *cada* contribuyente se beneficiarán *más* se lo que se habrían beneficiado si nadie hubiera contribuido. (2) Si contribuyen *menos* que k, los hijos de cualquier contribuyente se beneficiarán *menos* de lo que se habrían beneficiado si nadie hubiera contribuido. (1) y (2) hacen de k un umbral moral plausible por encima del cual cada padre o madre debe contribuir. Podemos afirmar

(R2) En tales casos, cada cual debe contribuir si cree que habrá al menos k contribuyentes.

Si nuestra Teoría del Acto Práctico afirma (R2), este cambio en nuestra concepción moral puede modificar con frecuencia lo que hacemos.

Puede decirse: «Como (1) es verdadero, no tenemos necesidad de afirmar que cada uno *debe* contribuir si cree que habrá al menos k contribuyentes. Como cada uno está componiendo una trama cuyo efecto neto es beneficiar a sus hijos, su amor por sus hijos hará que quiera componer esa trama. Hacerlo así será mejor para sus hijos».

Estas afirmaciones son falsas. Supongamos que al menos k padres contribuyen. Los hijos de estos contribuyentes se beneficiarán más de lo que se habrían beneficiado si nadie hubiera contribuido. Pero cada contribuyente está haciendo lo que es peor para sus propios hijos. Sería mejor para los hijos de cada contribuyente que él *no* contribuyera, y gastara todo lo que de este modo ahorrase —en tiempo, dinero, o energía— en beneficiar sólo a sus propios hijos. Esto es verdadero no importa cuántos de los otros contribuyan. Como cada contribuyente está haciendo lo que es peor para sus propios hijos, necesitamos afirmar que cada uno *debe* contribuir, si cree que habrá al menos k contribuyentes. Cada uno debe contribuir puesto que, aunque cada uno está haciendo lo que es peor para sus propios hijos, los k contribuyentes están haciendo *juntos* lo que es mejor para todos sus hijos.

Para apoyar (R2) podemos también señalar que, si algún padre o madre se queda sin contribuir cuando otros lo hacen, sus hijos irán de «polizones». Se beneficiarán de este bien público a expensas de los hijos de los contribuyentes. Se beneficiarán a sus expensas porque (a) se beneficiarán más que los hijos de los contribuyentes, y (b) esto es verdadero porque cada contribuyente hizo lo que era peor para sus propios hijos.

Afirmaciones parecidas se aplican al resto de obligaciones especiales que tenemos. Según la Moralidad del Sentido Común, debemos dar algún tipo de prioridad a los intereses de las personas con las que estamos relacionados de ciertas maneras. Además de nuestros hijos, algunos ejemplos son: nuestros padres, alumnos, pacientes, clientes, aquellos a los que representamos, y nuestros conciudadanos. Digamos que nosotros estamos *M-relacionados* con esas personas. Hay otras varias clases de relación M. Lo que estas relaciones tienen en común es que, según la Moralidad del Sentido Común, o M, tenemos obligaciones especiales con todos aquellos con los que estamos M-relacionados.

Hay innumerables casos en que, si cada cual diera prioridad a las personas con quienes está M-relacionado, esto sería, para todas estas personas, peor que si nadie les diera prioridad. Según (R1), lo que todos deberíamos idealmente hacer, en tales casos, es no dar ninguna prioridad a las personas con las que estamos M-relacionados. Si siguiéramos (R1), esto sería mejor para todas esas personas.

Supongamos que sabemos que algunos de nosotros darán prioridad a las personas con las que estamos M-relacionados. (R1) deja de aplicarse. Pero habrá de nuevo un número mínimo k tal que, si k o más *no* dan prioridad a las personas con las que están M-relacionados, esto será *mejor* incluso para *estas* personas de lo que habría sido si todos dieran prioridad a las personas con las que están M-relacionados. Podemos por consiguiente afirmar de manera plausible

> (R3) Cuando M es contraproducente, cada uno de nosotros *no* debe dar prioridad a aquellos con los que está M-relacionado, si considera que al menos k otros actuarán de la misma manera

Si aceptamos (R3), esto puede, de nuevo, cambiar a menudo lo que hacemos. Como antes, si cada uno de k personas no da prioridad a las personas con las que está M-relacionado, está haciendo lo

que es peor para esas personas. Por esta razón, de acuerdo con la Moralidad del Sentido Común, cada uno está obrando mal. Pero, aunque *cada uno* esté haciendo lo que es peor para las personas con las que está M-relacionado, estas k personas están *juntas* haciendo lo que es *mejor* para todas estas personas. Aunque estas k personas obren de un modo que según M está mal, juntas ocasionan que sean *mejor* logrados los fines M-dados de cada una. Como siguen (R3) en vez de M, juntas lo hacen mejor *incluso en términos de M*.

Consideremos a continuación las partes de nuestra teoría que dicen cuáles deben ser nuestros motivos. Supongamos que cada cual podría o bien (1) librar a su propio hijo de algún daño menor, o bien (2) librar al hijo de otro de algún daño mayor. De acuerdo con (R1), todos deberíamos idealmente hacer (2). Pero, ¿deberíamos estar *dispuestos* a hacer (2)? Si los daños menores, en sí mismos, fuesen grandes, tal disposición podría ser incompatible con el amor a nuestros hijos. Esto puede llevarnos a decidir que deberíamos seguir dispuestos a hacer (1). Si seguimos así dispuestos, podemos en tales casos, por tanto, hacer (1) en vez de (2). Nuestros hijos sufrirían entonces perjuicios mayores. Pero si los vamos a seguir queriendo, este es el precio que tenemos que pagar.

Vale la pena describir el caso extremo. Supongamos que tú y yo, cada uno de nosotros, tenemos cuatro hijos, y todos están en peligro mortal. No nos conocemos y no nos podemos comunicar. Cada uno podría o bien salvar a uno de sus propios hijos, o bien (2) salvar a tres de los hijos del otro. Si yo quiero a mis hijos, puedo encontrar imposible salvar las vidas de tres de tus hijos al precio de dejar morir a uno de los míos. Y lo mismo puede ser verdadero de ti. Ambos, entonces, haremos (1) en vez de (2). Puesto que queremos a nuestros hijos, salvamos a sólo dos de ellos, cuando podríamos haber salvado a seis.

Vistas las cosas en conjunto, será mejor si seguimos queriendo a nuestros hijos. Esto puede en ocasiones hacer causalmente imposible que hagamos lo que (R1) y (R2) afirman que debemos hacer. Pero habrá muchos otros casos en que esto no ocurriría. Así, es posible que ambos queramos a nuestros hijos y que contribuyamos a la mayoría de los bienes públicos.

Si volvemos a nuestras otras obligaciones especiales, es menos verosímil afirmar que deberíamos estar dispuestos a no hacer lo que (R3) afirma que debemos hacer. Así, los gobiernos de diferentes países deben poder no dar prioridad a sus propios ciudadanos, cuando esto fuese mejor para todos sus ciudadanos. Cuando consideramos los efectos de tener diferentes disposiciones, la concepción plausible, en la mayor parte de los casos, es que deberíamos estar dispuestos a actuar sobre la base de (R3).

He afirmado que, si debemos revisar la Moralidad del Sentido Común, debemos aceptar de la afirmación (R1) a la (R3). Como debemos querer a nuestros hijos, hay ciertos casos extremos en que no debemos estar dispuestos a actuar sobre la base de estas afirmaciones. Y puede haber otras excepciones similares. Pero en la mayor parte de los casos, debemos estar dispuestos a actuar a partir de ellas. Por eso deberíamos, a menudo, cambiar lo que hacemos.

39. Por qué debemos revisar la Moralidad del Sentido Común

Si revisamos la Moralidad del Sentido Común, o M, debemos aceptar tres de las afirmaciones de R: de (R1) a (R3). Vuelvo ahora a la cuestión principal. Si aceptamos M, ¿deberemos revisar nuestra concepción? ¿Deberemos pasar de M a R? ¿Es una objeción contra la Moralidad del Sentido Común el que, en muchos casos, sea contraproducente? Si lo es, R es el remedio obvio. R revisa M en el punto en que M es contraproducente. Y la única diferencia es que R no lo es.

Recordemos en primer lugar que, en estos casos, M es *directamente* contraproducente. El problema no es que, en nuestras tentativas de seguir M, fallemos de alguna manera. Eso simplemente haría a M indirectamente contraproducente. Como he mantenido, esto no plantearía ninguna objeción contra nuestra teoría. El problema es aquí más serio. En los casos que describí, todos seguíamos M *con éxito*. Cada uno tiene éxito al hacer lo que, de los actos que son posi-

bles para él, logra mejor sus fines M-dados. Pero, puesto que todos seguimos M con éxito, ocasionamos que los fines M-dados de cada uno sean *peor* logrados. Esto es lo que hace a M contraproducente. ¿Puede afirmarse que esto no constituye una objeción? Parece muy dudoso. Si hay alguna asunción sobre cuya base es clarísimo que una teoría moral *no* debería ser contraproducente es la asunción de que es universalmente seguida.

Recordemos además que por «fines» me refiero a fines *sustantivos*. He ignorado nuestro fin *formal*: la evitación de la maldad. Puede que parezca que esto elimina la objeción. Considérese esos casos en que, si seguimos M, o bien el resultado será peor para todos nuestros hijos, o bien cada uno podrá beneficiar menos a sus hijos. Podríamos decir: «Estos resultados son, desde luego, lamentables. Pero, ¿cómo podríamos evitarlos? Sólo fracasando en dar prioridad a nuestros propios hijos. Esto sería incorrecto. Así que estos casos no ponen en duda nuestra teoría moral. Aun para lograr nuestros otros fines morales, nunca obraríamos mal».

Estas observaciones no dan en el blanco. Es cierto que, en estos casos, M no es formalmente contraproducente. Si seguimos M, no estamos haciendo lo que creemos que está mal. Por el contrario, como creemos en M, creemos que está mal *no* seguir M. Pero M es sustancialmente contraproducente. A no ser que todos hagamos lo que ahora pensamos que es incorrecto, ocasionaremos que los fines M-dados de cada cual sean peor logrados. La cuestión es: ¿Podría esto mostrar que estamos equivocados? ¿Debemos hacer tal vez lo que *ahora pensamos* que es incorrecto? Podemos contestar, «No —nunca deberíamos obrar mal»—. Si estamos equivocados, *no* estaríamos obrando mal. Ni podemos decir, simplemente, «Pero, incluso en estos casos, *debemos* dar prioridad a nuestros propios hijos». Esto asume sólo que no estamos equivocados. Para defender nuestra teoría, tenemos que afirmar más que esto. Tenemos que afirmar que no representa objeción alguna contra ella el que, en tales casos, sea directa y sustantivamente contraproducente.

Esto no constituiría objeción alguna si simplemente no importara que nuestros fines M-dados vayan o no a ser logrados. Pero sí

que importa. El sentido en que importa puede necesitar explicación. Si no hemos obrado mal, puede que no importe moralmente. Pero importa en un sentido que tiene implicaciones morales. ¿Por qué deberíamos tratar de lograr nuestros fines M-dados? Parte de la razón es que, en este otro sentido, su logro importa. Si el logro de nuestros fines morales no importase, serían como un conjunto de reglas sin sentido, que tendrían como finalidad, simplemente, probar nuestra obediencia.

Ahora se puede decir: «Llamas a M contraproducente. Así que tu objeción tiene que apelar *a M*. No deberías apelar a ninguna teoría rival. Esto es lo que justamente has hecho. Cuando afirmas que importa el que nuestros fines M-dados sean logrados, estás afirmando meramente que, si no son logrados, el resultado será peor. Esto es asumir el Consecuencialismo. De modo que das por sentado lo que habría que probar».

Esto no es así. Al explicar por qué, combinaré una vez más dos distinciones. Cuando nuestros fines se mantienen en común, son neutrales respecto del agente. Otros fines son relativos al agente. Cualquier fin puede tener que ver o con lo que ocurre o con lo que hacemos. Esto nos da cuatro clases de fin. Se dan abajo cuatro ejemplos.

	Tienen que ver con lo que ocurre	Tienen que ver con lo que hacemos
neutrales respecto del agente	que los hijos no pasen hambre	que los hijos sean cuidados por sus propios padres
relativos al agente	que mis hijos no pasen, hambre	que yo cuide a mis hijos

Cuando afirmo que importa que nuestros fines M-dados se consigan, no estoy asumiendo que sólo importen los resultados. Algunos de nuestros fines M-dados tienen que ver con lo que nosotros *hacemos*. Así, el cuidado parental puede no ser para nosotros un simple medio. Ni estoy asumiendo ningún neutralismo respecto del agente. Como la Moralidad del Sentido Común es, la mayor parte de las veces, relativa al agente, esto sería dar por sentado lo que hay que probar. Pero no es lo que estoy haciendo.

Hay aquí dos puntos. Primero, no estoy asumiendo que lo que importa sea el logro *de fines M-dados*. Supongamos que yo pudiese o bien (1) promover mis propios fines M-dados, o bien (2) promover los tuyos más efectivamente. Según M, yo debería aquí hacer (1) más bien que (2). Con ello ocasionaría que los fines M-dados fuesen, en conjunto, peor logrados. Pero esto no hace a M contraproducente. Si sigo M, hago que *mis* fines M-dados sean *mejor* logrados. En mis ejemplos el asunto no es que, si todos nosotros hiciéramos (1) más bien que (2), haríamos que los fines M-dados fuesen peor logrados. El asunto es que nosotros ocasionamos que *cada uno de nuestros propios* fines M-dados sea peor logrado. Lo hacemos peor no sólo en términos neutrales respecto del agente sino también en términos relativos al agente.

El segundo punto es que esto puede tener importancia de un modo relativo al agente. Ayudará recordar la teoría del Propio Interés. En los Dilemas del Prisionero, esta teoría es directamente contraproducente. Si todos en vez de ninguno seguimos con éxito PI, ocasionaremos con ello que los fines PI-dados de cada uno sean peor logrados. Produciremos el resultado peor para cada uno. Si creemos en PI, ¿pensaremos que esto importa? ¿O importa sólo si cada uno logra su fin formal: la evitación de la irracionalidad? La respuesta está clara. PI da a cada uno el fin sustantivo de que el resultado sea, para él, lo mejor posible. El logro de este fin importa. E importa *de un modo relativo al agente*. Si creemos en PI, pensaremos que importa que, en los Dilemas del Prisionero, si todos seguimos PI, esto será peor para cada uno de nosotros. Aunque no refutan PI, estos casos son, en términos del propio interés, lamentables. Al afirmar esto, no necesitamos apelar a la forma de PI neutral respecto del agente: el Utilitarismo. La teoría del Propio Interés trata sobre la racionalidad más que sobre la moralidad. Pero la comparación muestra que, al discutir la Moralidad del Sentido Común, no necesitamos dar por sentado lo que hay que probar. Si importa que nuestros fines M-dados sean logrados, esto puede ser importante de un modo relativo al agente.

¿Esto importa? Nótese que no estoy preguntando si esto es *todo* lo que importa. No estoy sugiriendo que el logro de nuestro fin formal —la evitación de la maldad— sea un mero medio. Aunque algunos consecuencialistas lo asumen, no es lo que cree la mayoría de nosotros. Podemos incluso pensar que la evitación de la maldad siempre importa muchísimo. Pero esto es aquí irrelevante. Estamos preguntando si cuestiona M el que sea sustantivamente contraproducente. ¿Podría esto demostrar que, en tales casos, M es incorrecto? Puede ser cierto que lo que importa muchísimo sea que evitemos la maldad. Pero esta verdad no puede demostrar que M sea correcto. No puede ayudarnos a decidir lo que *es* malo.

¿Podemos afirmar que la evitación de la maldad es *todo* lo que importa? Si fuera así, mis ejemplos no enseñarían nada. Podríamos decir, «Ser sustantivamente contraproducente no es, en el caso de la Moralidad del Sentido Común, ser *irreparablemente* contraproducente». ¿Podemos defender nuestra teoría moral de este modo? En el caso de algunos fines M-dados tal vez podamos. Consideremos las promesas triviales. Podríamos creer tanto que deberíamos tratar de cumplir tales promesas, como que no importaría si, sin que sea culpa nuestra, fracasamos a la hora de hacerlo. Pero no tenemos tales creencias sobre todos nuestros fines M-dados. Si nuestros hijos sufren daños, o si podemos beneficiarles menos, esto tiene importancia. Nuestra moralidad *no* es un conjunto de reglas sin sentido, que tengan meramente la finalidad de probar nuestra obediencia.

Recordemos finalmente que, en mis ejemplos, M es colectiva *pero no individualmente* contraproducente. ¿Podría esto proporcionar una defensa? Esta es la cuestión central que he planteado. Es porque M es individualmente satisfactoria por lo que, a nivel colectivo, puede ser *directamente* contraproducente. ¿Por qué ocurre que, si todos hacemos (1) en vez de (2), seguimos M *satisfactoriamente*? Porque cada uno hace lo que, de los actos que son posibles para él, *mejor* logra sus fines M-dados. ¿No es tal vez objeción ninguna que *nosotros* ocasionemos con ello que los fines M-dados de cada uno sean *peor* logrados?

Ayudará de nuevo recordar la teoría del Propio Interés. En los Dilemas del Prisionero, PI es colectivamente contraproducente. Si

estuviéramos eligiendo un código colectivo, algo que vayamos a seguir todos, PI nos diría que la rechazáramos a ella misma. La elección guiada por el propio interés sería alguna versión de la moralidad. Pero los que creen en PI pueden afirmar que esto es irrelevante. Pueden decir: «La teoría del Propio Interés no afirma ser un código colectivo. Es una teoría de la racionalidad individual. Ser colectivamente contraproducente no es, en el caso de PI, ser irreparablemente contraproducente».

¿Podemos defender así la Moralidad del Sentido Común? Esto depende de nuestra concepción de la naturaleza de la moralidad y del razonamiento moral. Según la mayoría de las concepciones, la respuesta sería No. Según estas concepciones, la moralidad es esencialmente un código colectivo —una respuesta a la pregunta «¿Cómo deberíamos actuar *todos* nosotros?»—. Una respuesta aceptable a esta pregunta tiene que ser satisfactoria a nivel colectivo. La respuesta no puede ser directa y colectivamente contraproducente. Si creemos en la Moralidad del Sentido Común, deberíamos por ello revisar esta teoría para que no fuese contraproducente de esta forma. Deberíamos adoptar R.

Consideremos en primer lugar la concepción de Kant acerca de la naturaleza del razonamiento moral. Asumamos que estoy haciendo frente a uno de mis Dilemas de Padres. ¿Podría querer yo racionalmente que todos diesen prioridad a sus propios hijos, cuando esto sería peor para los hijos de todos, incluyendo los míos? La respuesta es No. Para los kantianos, la esencia de la moralidad es el paso de *cada uno* a *nosotros*. Cada cual debería hacer sólo aquello que pudiese racionalmente querer que todos hiciesen. Una moral kantiana no puede ser directa y colectivamente contraproducente.

Hay varios autores que aceptan una concepción kantiana de la naturaleza del razonamiento moral, pero que también aceptan alguna forma de la Moralidad del Sentido Común. Si mantienen su concepción kantiana, estos autores deben dar el paso a la versión revisada R.

Otros autores mantienen concepciones *constructivistas* de la naturaleza de la moralidad. Un código moral es, para ellos, algo que crea una sociedad, o aquello que sería racional para los miembros de una sociedad acordar que sea lo que gobierne su conducta. Esta es otra

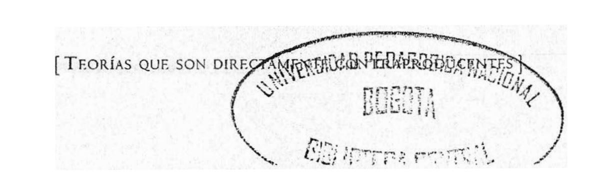
clase de concepción según la cual una teoría moral aceptable no puede ser directa y colectivamente contraproducente. Los que mantienen tal concepción no pueden seguir aceptando ninguna versión de la Moralidad del Sentido Común. Tienen que dar el paso a la correspondiente versión de R.

Los que mantienen una concepción constructivista pueden cuestionar mi división de la teoría moral. (R1) revisa lo que llamo nuestra Teoría del Acto Ideal. Los constructivistas puede que no vean ninguna necesidad para esta parte de una teoría moral. Pero no pueden objetar a mi propuesta que deberíamos preguntar lo que todos nosotros debemos hacer, simplemente según las asunciones de que todos lo intentaremos y todos tendremos éxito. Contestar a esta pregunta es en el peor de los casos innecesario. Si un constructivista pregunta lo que todos nosotros deberíamos idealmente hacer, su respuesta no puede ser ninguna versión de la Moralidad del Sentido Común. Si acepta alguna versión de esta moralidad, tiene que dar el paso a la correspondiente versión de (R1), la versión revisada de su moralidad que no sería directa y colectivamente contraproducente. Y, como debería aceptar (R1), debería aceptar también (R2) y (R3). Debería revisar su Teoría del Acto Práctico, la parte que solía ser toda su teoría.

Según la mayor parte de las otras concepciones de la naturaleza del razonamiento moral, la moralidad es esencialmente un código colectivo [58]. Según estas concepciones, si aceptamos la Moralidad del Sentido Común, tenemos que dar el paso a R. Pero algunos de los que creen en la Moralidad del Sentido Común puede que no tengan concepción ninguna de la naturaleza del razonamiento moral. ¿Podrían afirmar estas personas que, aunque haya muchos casos en que su código moral es directamente contraproducente, esto no plantea ninguna objeción al mismo, y no nos da ninguna razón para movernos a R?

Tal afirmación no es plausible. Y vale la pena sugerir cómo la Moralidad del Sentido Común está a punto de decirnos que demos

[58] Esto es cierto, por ejemplo, de los puntos de vista de Warnock, Mackie (2), Baier (1), Brandt (2), Harman, Gert, Toulmin, Rawls, y muchos otros.



el paso a R. Supongamos que, en uno de mis Dilemas de Padres, todos nosotros pudiéramos comunicarnos fácilmente. Entonces, la Moralidad del Sentido Común nos dirá que hagamos una promesa condicional conjunta en el sentido de que todos nosotros seguiremos, no este código moral, sino mi versión revisada R. Si me uno a otros en esta promesa condicional, será mejor para mis hijos. Mi obligación especial para con mis hijos se cumplirá por ello mejor si prometo condicionalmente, junto con todos los demás, que ninguno de nosotros dará prioridad a sus propios hijos. Hacer esta promesa condicional será lo mejor que yo pueda hacer por ellos. Si prometo seguir R, con la condición de que todos los demás prometan lo mismo, entonces los otros seguirán R sólo porque yo hice esta promesa. Si siguen R esto será mejor para mis hijos. Así que mi promesa produce un resultado mejor para ellos.

Observaciones similares se aplican a todos los otros casos en que la Moralidad del Sentido Común es contraproducente. Estos son casos en que creemos que debemos dar prioridad a las otras personas con las que estamos M-relacionados, tales como nuestros padres, alumnos, pacientes, clientes, o aquellos a los que representamos. En todos estos casos, si nos podemos comunicar fácilmente, la Moralidad del Sentido Común nos dirá que hagamos esta promesa condicional conjunta en el sentido de que seguiremos R. Entonces esto sería lo que cada cual debe hacer, a causa de lo que prometió. No estaríamos abandonando aquí la Moralidad del Sentido Común. Estaríamos usando parte de este código moral para alterar el contenido de lo que debemos hacer.

Supongamos, a renglón seguido, que en un Dilema de Padres, *no nos podemos* comunicar. Entonces seremos incapaces de lograr esta «solución moral». La Moralidad del Sentido Común nos dice ahora que cada uno dé prioridad a sus propios hijos. Esto será peor para todos nuestros hijos. Observaciones similares se aplican a los otros casos. Puesto que no nos podemos comunicar, y por tanto no podemos hacer la promesa conjunta, nuestro código moral no puede decirnos que sigamos R. Si *pudiéramos* comunicarnos, y revisar la pregunta, «¿Qué deberíamos hacer todos?», nuestra moralidad *nos diría* que prometiéramos *no* dar prioridad a nuestros propios hijos,

padres, alumnos, pacientes, etc. Nuestra moralidad nos diría que prometiéramos *no* hacer lo que, si no nos podemos comunicar, nos diría que hiciéramos. Está claro que, según nuestro código moral, sería *mejor* si pudiéramos comunicarnos, y pudiéramos entonces prometer seguir R. Esto aporta un sentido en que nuestra moralidad misma nos dice que aceptemos esta versión revisada de sí misma [59].

Hay una razón suplementaria para pensar que debemos revisar la Moralidad del Sentido Común. Esta teoría moral incurre en lo que llamé el Segundo Error en matemáticas morales. Ignora los efectos de conjuntos de actos —los efectos de lo que hacemos juntos—. El capítulo 3 demostró que esto es un error. Y, al demostrarlo, yo no estaba asumiendo el Consecuencialismo. Los que rechazan C estarían de acuerdo en que, en algunos de mis ejemplos, *no* debemos ignorar los efectos de lo que hacemos juntos.

La Moralidad del Sentido Común ignora estos efectos siempre que es directa y colectivamente contraproducente. Le dice a cada cual que haga lo que logre de la mejor manera posible sus fines M-dados. Esta afirmación asume que basta considerar los efectos de lo que hace cada persona. En estos casos, si cada uno hace lo que logra de la mejor manera posible sus fines M-dados, juntos ocasionamos que los fines M-dados *de cada uno* sean *peor* logrados. Esto es como un caso en que, si cada uno hace lo que no perjudica a nadie, juntos perjudicamos a muchas personas. En tales casos es un error pensar que lo que importa moralmente son sólo los efectos de lo que cada persona hace. Tenemos que estar de acuerdo en que

[59] (Nota añadida en 1987.) En su artículo en *Ethics*, julio de 1986, A. Kuflik sostiene que la Moralidad del Sentido Común no incluye la totalidad de lo que yo llamo M, y por eso no es contraproducente. Puede que sea así. Pero mi afirmación podría ser entonces que, si la Moralidad del Sentido Común *incluyera* la totalidad de M, *sería* contraproducente. Y lo que es más importante: a no ser que esta moralidad incluya mi revisión R, habrá muchos problemas de coordinación que, inaceptablemente, deja sin resolver. (Véase mi «Respuesta», en el mismo número.)

esto es un error, aunque rechacemos C y aceptemos la Moralidad del Sentido Común.

Supongamos que negamos que una teoría moral tenga que ser satisfactoria a nivel colectivo. Aunque neguemos esto, tenemos que admitir que, en los casos que he discutido, la Moralidad del Sentido Común incurre en un error, y debería por tanto ser revisada.

40. UNA REVISIÓN MÁS SIMPLE

Hay una revisión más simple de la Moralidad del Sentido Común, a favor de la que no he presentado argumentos. Esta es la forma completamente neutral en relación con el agente, o, para abreviar, *N*. Según esta teoría, cada uno de nosotros debería siempre tratar de hacer lo que, en conjunto, mejor logre los fines M-dados de todos. Nuestros fines morales relativos al agente se convierten en fines comunes.

Como está restringida a ciertos casos, la revisión R que propuse es complicada. N es simple, y teóricamente más atrayente. Esto sugiere que, si hemos dado el paso a R, deberíamos dar un paso más a N.

Puede parecer que podría ampliar el argumento de arriba para que diera esta conclusión. Consideremos un caso en que alguien podría o bien (1) promover sus propios fines M-dados, o bien (2) promover más efectivamente los fines M-dados de otros. Todos los casos así, tomados juntos, constituyen el *Caso que Todo lo Incluye*. Este incluye todos los casos en que la Moralidad del Sentido Común difiere de N. Si debemos revisar este código moral en este Caso que Todo lo Incluye, debemos aceptar N.

Supongamos que, en este Caso que Todo lo Incluye, todo el mundo hace (1) en vez de (2). Con ello ocasionaremos que nuestros fines M-dados sean, en conjunto, peor logrados. Si todos hiciéramos (2), serían, en conjunto, mejor logrados. Pero esto sería así sólo en conjunto, o para la mayoría de nosotros. Habría excepcio-

nes. No sería cierto que los fines M-dados *de cada uno* fueran a ser mejor logrados.

Recordemos el Dilema del Samaritano. Este es el cuarto ejemplo en la *Grundlegung** de Kant. ¿Podría querer yo racionalmente que fuese una ley universal que nadie ayudase a un desconocido en apuros? Para la mayoría de nosotros la respuesta es No. Pero hay algunas excepciones, como los ricos y los poderosos, que tienen guardaespaldas y sirvientes personales. Estas personas *podrían* querer racionalmente que nadie ayudara a los desconocidos en apuros. Sería peor para *casi* todos que nadie ayudara a esos desconocidos. Pero no sería peor para todos.

Una afirmación similar se aplica a mi Caso que Todo lo Incluye. Si todos siguiéramos satisfactoriamente la forma de M que es neutral en relación con el agente, sería verdadero de la mayoría de las personas que *sus propios* fines M-dados serían mejor logrados. Pero esto no sería verdadero de todo el mundo. Según la definición que di, no es verdadero *para todo el mundo* que M sea aquí directa y colectivamente contraproducente. (Para poner este punto de otra manera. En esa definición, «todo el mundo» significa «todos los miembros de algún grupo». En el Caso que Todo lo Incluye, la mayoría de las personas estarían en este grupo. Pero habría algunas ajenas a él).

Al aplicar el Test Kantiano, quizás podamos insistir en que el rico y el poderoso corran un velo de ignorancia. Puede decirse que este es uno de los requisitos del razonamiento moral. Pero no puedo hacer la afirmación correspondiente con el Caso que Todo lo Incluye. Mi argumento se dirige, no a los pre-morales para introducirlos en la moralidad, sino a los que creen en la Moralidad del Sentido Común. Como estas personas ya mantienen una concepción moral, no puedo afirmar similarmente que tienen que correr un velo de ignorancia. Por eso no he usado este argumento para la forma completamente neutral con relación al agente de la Moralidad del Sentido Común. He argumentado sólo

* Se refiere Parfit a la *Fundamentación de la Metafísica de las Costumbres*, publicada por Kant en 1785. (N. del t.)

a favor de la versión R más restringida. R se aplica sólo a esos casos en que, si todos seguimos M en vez de R, con ello ocasionaremos que los fines M-dados *de cada uno* sean peor logrados. Todos lo haremos peor no sólo en términos neutrales respecto del agente sino también en términos relativos al agente. Esta es una diferencia crucial. En el Caso que Todo lo Incluye, puedo afirmar sólo que, si todos seguimos M, con ello ocasionaremos que nuestros fines M-dados, en conjunto, sean peor logrados. Afirmar que esto importa es asumir el Neutralismo respecto del agente. Esta afirmación no puede ser parte de un argumento a favor del Neutralismo respecto del agente, puesto que esto *daría* por sentado lo que hay que probar.

5

CONCLUSIONES

En la primera parte de este libro he preguntado por lo que se revela cuando una teoría es contraproducente. Pasemos revista brevemente a las respuestas.

41. Reduciendo la distancia entre M y C

En los Dilemas del Prisionero, la teoría del Propio Interés es directa y colectivamente contraproducente. En estos casos, si todos buscamos el propio interés, esto será peor para todos nosotros. Sería mejor para todos nosotros si, en cambio, todos actuamos moralmente. Algunos pensadores argumentan que, como esto es verdadero, la moral es superior a la teoría del Propio Interés, incluso en términos del propio interés.

Como mostré en el capítulo 4, este argumento falla. En estos casos, PI triunfa a nivel individual. Como PI es una teoría de la racionalidad individual, no necesita ser satisfactoria a nivel colectivo.

Cuando este argumento lo defienden los que creen en la Moralidad del Sentido Común, *les sale el tiro por la culata*. No refuta a PI, pero *sí* refuta parte de su propia teoría. Como PI, la Moralidad

del Sentido Común es con frecuencia directa y colectivamente contraproducente. A diferencia de PI, una teoría moral tiene que ser colectivamente satisfactoria. Los que creen en M tienen por tanto que revisar sus creencias, moviéndose de M a R. Su Teoría del Acto Ideal debería incluir (R1), y su Teoría del Acto Práctico debería incluir (R2) y (R3).

A diferencia de M, R es consecuencialista, dándonos a todos nosotros fines morales comunes. Como el capítulo 4 muestra que los que creen en M tienen que dar el paso a R, esto reduce el desacuerdo entre la Moralidad del Sentido Común y el Consecuencialismo.

El capítulo 1 también reduce este desacuerdo. Somos puros bienhechores si siempre tratamos directamente de seguir C, haciendo todo lo que produzca el mejor resultado de todos. C es indirecta y colectivamente contraproducente. Si todos fuéramos puros bienhechores, el resultado sería peor de lo que sería si tuviéramos ciertos otros deseos y disposiciones. Este hecho no refuta C, pero muestra que C tiene que incluir Teorías del Motivo Ideal y Práctico. La Teoría del Motivo Ideal de C tiene que afirmar que no deberíamos ser todos puros bienhechores. La Teoría del Motivo Práctico de C tiene que afirmar que cada uno de nosotros debería tratar de tener uno de los mejores conjuntos de deseos y disposiciones posibles, en términos de C. (Alguien tiene uno de estos conjuntos si no hay otro posible conjunto del cual sea cierto que, si esta persona tuviera este otro conjunto, las consecuencias serían mejores.)

Para la mayoría de nosotros, las mejores disposiciones *corresponderían* en líneas generales a la Moralidad del Sentido Común en el siguiente sentido. A menudo deberíamos estar fuertemente dispuestos a hacer lo que esta moralidad requiere.

Aquí tenemos dos de las maneras en las que esto es así. Producirá un resultado mejor el que la mayoría de nosotros tenga los intensos deseos de los que la mayor parte de la felicidad depende. Así que será mejor que amemos a nuestras familias y a nuestros amigos. Entonces estaremos fuertemente dispuestos a dar ciertos tipos de prioridad a los intereses de personas como nuestros alumnos, pacientes, protegidos, clientes, o aquellos a los que representamos. Actuando de esta manera estaríamos haciendo de nuevo lo que M afirma que debemos hacer.

Por diferentes razones, deberíamos estar fuertemente dispuestos a *no* actuar de cierto modo. Si lo que queremos es que alguien esté muerto, probablemente llegaremos a creer, de modo falso, que la muerte de esta persona produciría un resultado mejor. Por tanto, deberíamos estar fuertemente dispuestos a no matar a los demás. Afirmaciones similares se aplican a engañar o coaccionar a otros, ceder a las amenazas, y a otras clases de actos de los que la Moralidad del Sentido Común afirma que son incorrectos.

42. Hacia una Teoría Unificada

Puesto que C es una teoría neutral respecto del agente, es indirectamente contraproducente, y por ello necesita incluir Teorías del Motivo Práctico e Ideal que, en el sentido definido arriba, correspondan en líneas generales a M. Puesto que M es una teoría relativa al agente, es con frecuencia directamente contraproducente, y por eso necesita ser revisada para que sus Teorías del Acto Práctico e Ideal sean en parte consecuencialistas. C y M se enfrentan a objeciones que pueden ser anuladas únicamente ampliando y revisando estas teorías, de maneras que las aproximen la una a la otra.

Estos hechos sugieren naturalmente una posibilidad atractiva. Los argumentos de los capítulos 1 y 4 dan apoyo a conclusiones que pueden *encajar*, o unirse para construir un todo mayor. Podríamos ser capaces de desarrollar una teoría que incluyera y combinara versiones revisadas tanto de C como de M. Llamémosla la *Teoría Unificada*.

Estas afirmaciones son como las que hicieron Sidgwick, Hare y otros [60]. Pero hay al menos dos diferencias:

(1) La mayoría de estos pensadores tratan de combinar la Moralidad del Sentido Común y el Utilitarismo, o *U*.

[60] Véase Sidgwick (1), Libros III y IV (y el comentario en Schneewind), Hare (1), Sumner (1), y Sumner (2), capt. 5.

Sidgwick argumenta a favor de la versión hedonista de U; Hare argumenta a favor de la versión de la realización de deseos. Yo he estado discutiendo la teoría más amplia, el Consecuencialismo. C puede apelar a diversos principios sobre lo que hace malos a los resultados. C puede afirmar, por ejemplo, que sería peor si hubiera más desigualdad, engaño y coerción, y no se respetaran ni se cumplieran los derechos de las personas. Si C hace estas afirmaciones, C está ya, comparada con U, más próxima a la Moralidad del Sentido Común.

(2) En mi afirmación de que C es indirectamente contraproducente, simplemente estoy siguiendo a Sidgwick y a Hare. Pero no sigo a ninguno de los dos en mi argumento contra la Moralidad del Sentido Común. Este argumento, a diferencia del de Hare, no apela a una teoría particular sobre la naturaleza de la moral, o a la lógica del lenguaje moral. Y mi argumento, a diferencia del de Sidgwick, no apela a nuestras intuiciones. Yo afirmo que la Moralidad del Sentido Común es en muchos casos directa y colectivamente contraproducente. Esta afirmación no requiere de ninguna asunción aparte de las que se hacen por la Moralidad del Sentido Común. Cuando concluyo que, en estos casos, esta moralidad tiene que ser revisada, asumo que una teoría moral tiene que ser satisfactoria a nivel colectivo. Pero esta asunción no se hace de parte de una única teoría sobre la naturaleza de la moralidad. O bien se hace o bien viene implicada por la mayoría de las muy diversas teorías acerca de este tema. Y sería aceptada por la mayoría de los que creen en alguna versión de la Moralidad del Sentido Común.

Los dos argumentos de los capítulos 1 y 4 señalan a una Teoría Unificada. Pero desarrollar esta teoría de forma convincente llevaría al menos un libro. Ese libro no es este libro. Me limitaré a añadir algunos comentarios breves.

43. Trabajo por hacer

De acuerdo con C, a menudo deberíamos estar fuertemente dispuestos a actuar de la forma que M requiere. Si creemos en C y no sólo tenemos sino que además actuamos sobre la base de estas disposiciones, los que creen en M no pueden objetar nada a lo que nosotros *hacemos*. Pero pueden plantear objeciones a nuestras *creencias*. Como nuestras disposiciones van a actuar de maneras que a menudo harán el resultado peor, con frecuencia creeríamos que actuando sobre la base de esas disposiciones, como los creyentes en M piensan que deberíamos hacer, estaríamos actuando mal. Al desarrollar una Teoría Unificada, nuestra tarea mayor sería la de reconciliar estas creencias en conflicto.

Además de afirmar que deberíamos tener estas disposiciones, C puede afirmar también que debería ser nuestra *política* actuar de acuerdo con ellas. C puede afirmar que esta debería ser nuestra política, aunque estas no sean disposiciones a hacer lo que produciría el mejor resultado de todos. Puede ser cierto, en los modos descritos en el capítulo 1, que seguir esta política produciría el mejor resultado de todos. Recordemos a continuación la Teoría de la Reacción de C. Esta afirma que debemos sentir remordimiento y culpar a otros cuando esto produzca el mejor resultado de todos. Si seguimos la política que acabamos de describir, a menudo dejaremos de hacer lo que produciría el mejor resultado de todos. C afirmará, por tanto, que, en un sentido, estamos obrando mal. Pero C puede también afirmar que, puesto que estamos siguiendo esta política, no deberíamos ser culpados, ni sentir remordimiento. C podría implicar que deberíamos ser culpados y sentir remordimiento sólo si *no* seguimos esa política. Esto podría ser la pauta de culpa y remordimiento que produciría el mejor resultado de todos. Si C hiciera estas afirmaciones, esto reduciría el conflicto entre C y M. Aunque estas teorías todavía estarían en desacuerdo en lo que respecta a qué actos son correctos o incorrectos, estarían en desacuerdo mucho menos en lo que respecta a qué actos son dignos de censura, y cuáles deberían generar en nosotros remordimiento. Estaríamos más cerca de la Teoría Unificada.

Estas últimas afirmaciones están enormemente simplificadas. Al desarrollar la Teoría Unificada, necesitaríamos tanto considerar muy diferentes tipos de actos y de políticas, como considerar de qué manera estos estarían relacionados con cosas como nuestras emociones, necesidades y habilidades. Habría que responder muchas

preguntas. Para ser convincente, la Teoría Unificada tiene que trazar muchas distinciones y hacer muchas afirmaciones diferentes. Habría mucho trabajo que hacer, y yo no voy a intentar hacerlo aquí.

Como la Teoría Unificada incluiría una versión de C, puede objetarse que sería demasiado exigente. Pero esta objeción puede también anularse parcialmente con la Teoría de la Reacción de C.

Volvamos a la cuestión de cuánto deberían dar a los pobres los de las naciones ricas. Como los otros, en efecto, darán poco, C afirma que cada rico debe dar casi todos sus ingresos. Si los ricos dan menos, están obrando mal. Pero si se culpase a cada rico por dejar de dar casi todos sus ingresos, la censura dejaría de ser efectiva. La mejor pauta de culpa y remordimiento es la que ocasionara que los ricos dieran más. Puestas así las cosas, C podría implicar que los ricos deberían ser culpados y deberían sentir remordimiento sólo cuando dejaran de dar una parte muy pequeña de sus ingresos, como por ejemplo un décimo.

Comparada con la Moralidad del Sentido Común, C es de otros modos mucho más exigente. Por ejemplo, C afirmaría a veces que uno debería sacrificar si vida para que se pudieran salvar unos desconocidos. Pero no salvar a estos desconocidos no sería, incluso en los términos de C, censurable. Como incluiría a C, la Teoría Unificada sería más exigente que la Moralidad del Sentido Común, como esta es ahora. Pero si hace estas afirmaciones sobre la culpa y el remordimiento, sus demandas puede que no sean razonables o realistas.

44. Otra posibilidad

Muchas personas son *escépticos morales*: creen que no puede ser verdadera ninguna teoría moral, o que no hay ninguna que sea la mejor. Puede ser difícil resistirse al escepticismo si seguimos teniendo desacuerdos profundos. Uno de nuestros más profundos desacuerdos es el que se produce entre los consecuencialistas y los que creen en la Moralidad del Sentido Común.

Los argumentos en los capítulos 1 y 4 reducen este desacuerdo. Si podemos desarrollar la Teoría Unificada, incluso se podría eliminar. Podríamos descubrir que, para decirlo en palabras de Mill, nuestros oponentes estaban «escalando la colina por el otro lado».

Puesto que nuestras creencias morales ya no estarían en desacuerdo, también podríamos cambiar nuestra concepción acerca del status de las mismas. El escepticismo moral podría ser socavado.

SEGUNDA PARTE

RACIONALIDAD Y TIEMPO

6

LA MEJOR OBJECIÓN A LA TEORÍA DEL PROPIO INTERÉS

45. La teoría del fin Presente

Los argumentos de la Primera Parte no refutaron la teoría del Propio Interés. Ahora presentaré otros argumentos en su contra. Algunos de ellos apelan a la moralidad. Pero el desafío principal viene de una teoría diferente de la racionalidad. Esta es la teoría del fin Presente, o P.

Describí tres versiones de P. Una es la *teoría Instrumental*. Esta afirma

> *PIn*: Lo que cada uno de nosotros tiene más razón para hacer es lo que realizaría mejor sus deseos presentes.

Amplío la palabra «deseo» para que incluya intenciones, proyectos y otros fines.

Para aplicar la teoría Instrumental, tenemos que ser capaces no sólo de distinguir deseos diferentes, sino también de decidir qué deseos tiene uno realmente. Las dos cosas pueden ser difíciles. Pero aquí ignoraré estas dificultades. Todo lo que me hace falta asumir es que a veces podemos decidir lo que, en conjunto, mejor realizaría los deseos presentes de alguien.

Al decidir esto, deberíamos ignorar deseos *derivados*. Se trata de deseos de lo que son meros medios para la realización de otros deseos. Supongamos que quiero ir a cierta biblioteca simplemente para poder encontrarme con una hermosa bibliotecaria. Si tú me presentas a esta bibliotecaria, no me queda ningún deseo sin realizar. Por «deseos» significaré «deseos no derivados».

Al decidir lo que mejor realizaría estos deseos, deberíamos asignar mayor peso a los que son más intensos. El deseo más intenso de una persona puede ser contrarrestado por otros deseos diversos. Supongamos que, si yo hiciera X, esto *no* realizaría mi deseo presente más intenso, pero satisfaría *todos* mis otros deseos presentes. Aunque X no realizara mi deseo más intenso, yo puedo decidir que, de los actos que son posibles para mí, X es lo que en conjunto mejor satisfaría mis deseos. Si decido esto, X puede convertirse en lo que, una vez consideradas todas las cosas, yo deseo hacer en mayor medida.

En su tratamiento del conflicto entre deseos, y de las decisiones en situaciones de riesgo e incertidumbre, la teoría Instrumental puede adoptar formas sutiles. Sin embargo, como su nombre indica, se halla totalmente interesada en la elección de medios. No somete a crítica los *fines* del agente —*lo que* él desea—. Como es bien sabido, Hume escribió: «No es contrario a la razón preferir la destrucción del mundo entero a hacerme un rasguño en mi dedo. No es contrario a la razón elegir mi ruina absoluta para evitarle la más mínima incomodidad a un indio, o a una persona que me resulte totalmente desconocida. No es tampoco contrario a la razón preferir incluso un bien pequeño, aun reconociéndolo menor, a otro mayor...» [1].

Esta negativa a criticar deseos no es una parte esencial de la teoría del fin Presente. Hasta Hume sugirió que «una pasión... puede ser llamada "irrazonable"... cuando se funda en una suposición falsa». Esta sugerencia se desarrolla en la *Teoría Deliberativa*. Esta afirma

> *PD*: Lo que cada uno de nosotros tiene la máxima razón para hacer es aquello que lograría mejor, no lo que cada uno *realmente* quiere, sino lo que cada uno *querría*, en el momento de actuar, si hubiera pasado por un proceso de «deliberación ideal» —si conociera los hechos relevantes, pensara con claridad y estuviera libre de influencias distorsionantes.

[1] Hume (1), Libro II, Parte III, Sección III.

Los hechos relevantes son esos de los que es verdadero que, si esta persona conociera este hecho, sus deseos cambiarían. Esta última afirmación necesita refinarse, de modos que aquí podemos ignorar [2].

Una tercera versión de P es la *teoría Crítica del fin Presente*, o *CP*. Esta afirma que ciertos deseos son intrínsecamente irracionales, y no proporcionan buenas razones para actuar. CP puede afirmar también que ciertos deseos están racionalmente requeridos. Según esta segunda afirmación, una persona que no tuviera estos deseos sería irracional.

Tenemos que distinguir aquí entre dos clases de razón: *explicativa* y *buena*. Si alguien obra de cierto modo, podemos saber cuál fue su razón. Al describir esta razón, explicamos por qué esta persona obró como lo hizo. Pero puede que creamos que esta razón era una razón muy mala. Por «razón» me referiré a la «buena razón». Según este uso, afirmaríamos que esta persona *no* tenía razón para obrar como lo hizo.

Según la teoría Deliberativa, *cualquier* deseo proporciona una razón para actuar, con tal de que sobreviva al proceso de deliberación racional. Supongamos que, conociendo los hechos y pensando con claridad, yo prefiero la destrucción del mundo a arañarme mi dedo. Según la teoría Deliberativa, si tuviese una Máquina del Día del Juicio Final, y pudiera actuar a partir de esta preferencia, sería racional que lo hiciera así. Podemos rechazar esta afirmación. Puede que creamos que *esta* preferencia no proporciona una razón para actuar. Y puede que creamos lo mismo en relación con muchos otros deseos posibles. Según la teoría Deliberativa, no hay deseo que sea intrínsecamente irracional. Si creemos que tales deseos existen, entonces deberíamos rechazar esta teoría.

[2] La Teoría Deliberativa se desarrolla del modo más completo en Brandt (2).

Un teórico deliberativo podría responder que esos deseos irracionales no sobrevivirían al proceso de deliberación ideal. Esto podría hacerse trivialmente verdadero con una definición. El teórico podría decir que nadie que tenga estos deseos puede «pensar con claridad». Pero esto, en efecto, lo que hace es conceder la objeción. Al definir lo que cuenta como «pensar con claridad», el teórico tendría que referirse a los deseos en cuestión. Tendría que decidir qué deseos son intrínsecamente irracionales.

El teórico deliberativo podría, en cambio, hacer de su respuesta una afirmación fáctica. Podría estar de acuerdo en que lo que quiere decir con «pensar con claridad» no excluye tener los deseos que creemos intrínsecamente irracionales. Pero podría insistir en que su teoría es adecuada, puesto que los que pensaran con claridad y conocieran los hechos no tendrían tales deseos.

Si esto es así resulta difícil de predecir. Y aunque fuese así, nuestra objeción no habría sido anulada del todo. Si ciertas clases de deseo son intrínsecamente irracionales, cualquier teoría de la racionalidad que sea completa debe afirmarlo. No deberíamos ignorar la cuestión de si hay tales deseos simplemente porque tenemos la esperanza de que, si pensamos con claridad, nunca los tendremos. Si pensamos que puede haber tales deseos, deberíamos desplazarnos desde la versión deliberativa a la versión Crítica de la teoría del fin Presente.

Esta teoría, curiosamente, ha sido descuidada. Las versiones instrumental y deliberativa, en las que se cree ampliamente, hacen dos afirmaciones: (1) Lo que cada persona tiene más razón para hacer es lo que realizaría mejor los deseos que, en el momento de actuar, dicha persona tiene o tendría si conociera los hechos y pensara con claridad. (2) Los deseos no pueden ser intrínsecamente irracionales, ni venir requeridos racionalmente. Estas son afirmaciones muy diferentes. Podríamos rechazar (2) y aceptar una versión matizada de (1). Entonces aceptaríamos la versión Crítica de P. Esta afirma

> CP: Ciertos deseos son intrínsecamente irracionales. Y un conjunto de deseos puede ser irracional aunque los deseos de este conjunto no sean irracionales. Por ejemplo, es irracional preferir X a Y, Y a Z y Z a X. Un conjunto de deseos puede también ser irracional porque falla en contener deseos que vienen requeridos racionalmente. Supongamos que conozco los hechos y estoy pensando con claridad. Si mi conjunto de deseos no es irracional, lo que tengo más razón para hacer es lo que realizaría mejor aquellos de mis deseos presentes que no son irracionales. Esta afirmación se aplica a todos en cualquier tiempo.

La acusación de «irracional» se encuentra en un extremo de una gama de críticas. Es como la acusación de «malvado». Podemos afirmar que determinado acto, aunque no tan malo como para ser malvado, está aún así abierto a la crítica moral. Podemos afirmar de manera parecida que determinado deseo, aunque no merece la acusación extrema de «irracional», está abierto a la crítica racional. Para ahorrar palabras, ampliaré el uso corriente de «irracional». Usaré esta palabra para significar «abierto a la crítica racional». Esto permitirá que «no irracional» signifique «no abierto a tal crítica».

En su afirmación sobre los deseos que no son irracionales, CP no necesita apelar únicamente a la intensidad de estos deseos. Puede, por ejemplo, no asignar peso a esos deseos que alguien desea no tener. Y CP no necesita apelar únicamente a lo que, aún en el sentido más amplio, podemos llamar deseos. Puede también apelar a los valores, ideales y creencias morales del agente. Todas estas cosas pueden proporcionar razones para actuar. Pero CP afirma que algunas de estas no son buenas razones. Puede afirmar, por ejemplo, que incluso para las personas que creen en la etiqueta, o en algún código de honor, no hay razón para obedecer a ciertas reglas sin sentido, o para luchar en los duelos que el honor exija.

He descrito tres versiones de la teoría del fin Presente. Aunque esta descripción es un simple boceto, aquí será suficiente. Mucho de lo que sigue se centrará en lo que estas diferentes versiones tienen en común. En parte por esta razón, sólo voy a discutir casos en que las teorías Deliberativa e Instrumental coinciden. Hay casos en que determinada persona conoce todos los hechos relevantes, y además piensa con claridad. Asumiré también que lo que mejor realizaría los deseos presentes de esta persona es lo mismo que lo que quiere en mayor medida, después de considerar todas las cosas. Y

con frecuencia asumiré que los deseos de esta persona no entran en conflicto ni con sus creencias morales ni con sus otros valores e ideales. Al hacer estas asunciones evito considerar varias cuestiones importantes. Estas cuestiones tendrían que ser resueltas por cualquier teoría completa de la racionalidad. Pero no son relevantes para mi objetivo principal en la Segunda Parte de este libro. Este objetivo es mostrar que deberíamos rechazar la teoría del Propio Interés.

Puesto que este es mi objetivo principal, la Segunda Parte puede resultar aburrida a los que ya rechazan esta teoría. Pero discutiré también algunas cuestiones desconcertantes acerca de la racionalidad y el tiempo. Y apoyaré CP, afirmando que ciertos deseos son intrínsecamente irracionales, y que otros pueden venir racionalmente requeridos. Puesto que estas afirmaciones están sometidas a controversia, les dedicaré una defensa antes de volver a la teoría del Propio Interés.

46. ¿PUEDEN SER LOS DESEOS INTRÍNSECAMENTE IRRACIONALES, O VENIR RACIONALMENTE REQUERIDOS?

¿Por qué piensa la gente que un deseo no puede ser irracional a no ser que se apoye en una creencia falsa? Podemos recordar para empezar por qué Hume, su más distinguido defensor, defendía esta concepción. Hume entendió que el razonamiento versaba solamente sobre creencias, y sobre la verdad y la falsedad. Un deseo no puede ser falso. Y un deseo puede ser «denominado irrazonable», según el modo de ver de Hume, sólo si conlleva irracionalidad teórica.

El razonamiento *no* tiene que ver sólo con creencias. Además de razones para creer, hay razones para actuar. Además de racionalidad teórica hay racionalidad práctica. Hay así un modo diferente y más simple en que un deseo puede ser irracional. Puede ser un deseo que no proporciona una razón para actuar.

Algunos seguidores de Hume se resisten a llamar a los deseos «irracionales». Esta diferencia sería trivial si estuvieran de acuerdo en que determinados deseos no proporcionan buenas razones para

actuar. Recordemos a continuación que uso «irracional» para referirme a «expuesto a la crítica racional». Se trata de una cuestión de grado. Si un deseo no es completamente irracional, puede proporcionar una razón para actuar. Si uno de dos deseos se halla más expuesto a la crítica racional, proporciona una razón más débil.

Si un deseo es completamente irracional, no proporciona *directamente* razón alguna para actuar. Pero hay deseos que, aunque irracionales, proporcionan *indirectamente* tales razones. Si sufro de claustrofobia, puede que tenga un intenso deseo de no estar en un espacio cerrado. Este deseo es como el miedo. Dado que el miedo conlleva la creencia de que el objeto temido es peligroso, el miedo es irracional cuando esta creencia es claramente falsa. Supongamos que, cuando tengo ese intenso deseo de huir de un espacio cerrado, sé que en realidad no estoy en peligro. Ya que mi deseo es como el miedo, puedo juzgarlo irracional. Pero este deseo proporciona indirectamente una razón para actuar. Hace que le tenga una intensa aversión a estar en ese espacio cerrado, y tengo una razón para tratar de escapar a lo que me disgusta tan intensamente.

Cuando un deseo proporciona directamente una razón para actuar, la razón rara vez es el deseo. Rara vez ocurre que, cuando alguien actúa de algún modo, su razón simplemente sea que quiere hacerlo. En la mayoría de los casos, la razón que tiene alguien para actuar es uno de los rasgos de lo que quiere, o uno de los hechos que explica y justifica su deseo. Supongamos que ayudo a alguien que lo necesita. Mi razón para ayudar a esa persona no es que yo la quiera ayudar, sino que ella necesita ayuda, o que yo le prometí ayudarla, o algo por el estilo. De forma similar, mi razón para leer un libro no es que yo lo quiera leer, sino que el libro es ingenioso, o que explica por qué existe el universo, o que de algún otro modo vale la pena leerlo. En ambos casos, mi razón no es mi deseo sino el aspecto en el cual lo que estoy haciendo vale la pena hacerse, o el aspecto en el cual mi fin es *deseable* —digno de ser deseado [3].

El que una razón casi nunca sea un deseo puede parecer que socava la teoría del fin Presente. Puede parecer que revela que lo que

[3] Sigo a Anscombe y a Norman.

tenemos más razón para hacer no puede depender de lo que deseamos. Pero esto es falso. Aunque una razón no sea un deseo, puede depender de un deseo. Supongamos que mi razón para leer un libro es que explica las causas de la Primera Guerra Mundial. Si no tuviera ningún deseo de saber cuáles fueron estas causas, no tendría razón alguna para leer este libro.

Afirmé que, de acuerdo con CP, hay deseos que pueden venir requeridos racionalmente. Volvamos al caso en que mi razón para ayudar a alguien es que necesita ayuda. ¿Esta razón depende de un deseo? ¿Tendría yo una razón para ayudar a esta persona aunque no me importaran sus necesidades? Más en general, ¿tendría yo una razón para actuar moralmente aunque no me importara la moralidad?

Estas son dos preguntas sometidas a controversia. Algunos autores las contestan con un No. Según ellos, estas dos razones dependen esencialmente de mis deseos, o de lo que a mí me importa [4].

Otros autores afirman que tengo una razón para actuar moralmente, o para ayudar a quien lo necesite, aunque yo no tenga ningún deseo de actuar así. Esta afirmación entra en conflicto con la versión Instrumental de P, y puede entrar en conflicto con la versión Deliberativa. Pero no es necesario que entre en conflicto con la versión Crítica. Si aceptamos CP, podríamos afirmar

> (CP1) Cada uno de nosotros está racionalmente requerido tanto a preocuparse por la moralidad como a preocuparse por las necesidades de los demás. Puesto que esto es así, tenemos una razón para actuar moralmente, aunque no tengamos ningún deseo de hacerlo. Si tenemos una razón para actuar de cierto modo depende usualmente de si tenemos ciertos deseos. Pero esto no es así en el caso de los deseos que son requeridos racionalmente.

Esto lo *podría* decir todo el que aceptara CP. Mi descripción de CP deja abierta la cuestión de si esto *debería* decirse. Como se trata

[4] Véase Foot, y Williams (5).

de una cuestión controvertida, es mejor dejarla abierta. Cualquiera de las dos respuestas las podría dar el que aceptara CP. Si una teoría dejara abierta *toda* cuestión controvertida, no valdría la pena discutirla. Pero, como ya veremos, esto no ocurre con CP.

Otra afirmación distintiva de CP es que hay deseos, o conjuntos de deseos, que son intrínsecamente irracionales. Escribí más arriba que, en la mayor parte de los casos, mi razón para actuar no es uno de mis deseos, sino el aspecto en que lo que deseo vale la pena de ser deseado. Esto sugiere de forma natural cómo hay deseos que podrían ser intrínsecamente irracionales. Podemos afirmar: «Es irracional desear algo que no merece ser deseado en ningún aspecto. Es aún más irracional desear algo que merece *no* ser deseado —que merece ser evitado».

Se podría decir que los masoquistas tienen deseos así. Pero el masoquismo genuino es un fenómeno complicado que necesitaría una larga discusión. Podríamos imaginar un caso más simple, en el que alguien simplemente quiere, en un futuro indeterminado, sufrir gran dolor. Supongamos que, a diferencia de los masoquistas, esta persona sabe que no va a disfrutar en absoluto de este dolor, y tampoco va a encontrar que disminuye su sentimiento de culpabilidad, ni va a ser beneficiada por el dolor de ninguna otra manera. Esta persona simplemente quiere tener sensaciones que, cuando llegue el momento, aborrecerá intensamente, y querrá con mucha fuerza no tener en absoluto. La mayoría de nosotros creería que este deseo es irracional.

¿Hay casos reales de personas que tienen deseos irracionales? Un ejemplo puede ser el deseo de saltar que tienen algunos cuando están al borde de un precipicio. Este extraño impulso lo experimentan personas que no tienen el más mínimo deseo de morir. Como estas personas quieren seguir viviendo, puede ser irracional para ellas *actuar de acuerdo con* su deseo de saltar. Pero esto no muestra que su deseo sea irracional. El deseo de saltar *no* es un deseo de morir. En el caso de algunos, es un deseo de subir por el aire. Lo que es algo que merece desearse; podemos envidiar a los pájaros de una manera racional. Hay otros que aunque quieren saltar no tienen

el más mínimo deseo de subir por el aire. En su caso, su deseo de saltar puede ser irracional.

Consideremos a continuación otra forma que adopta este deseo. Se dice que, en la cumbre del éxtasis, algunas parejas japonesas se tiran de un salto al precipicio, porque *realmente* quieren morir. Quieren morir *porque* han llegado a la cumbre del éxtasis. ¿Acaso puede *esto* ser una razón para suicidarse?

Unos dirían: «No. Vale la pena desear la muerte cuando vaya a poner un final a nuestra agonía. Pero no vale la pena desearla porque vaya a poner un final a nuestro éxtasis».

Esto describe mal el caso. Estas parejas no quieren morir porque esto vaya a poner un final a su éxtasis. Quieren que su vida termine en el punto más elevado, en el mejor punto. No es esto lo que la mayoría de nosotros queremos. Ahora bien, aunque este deseo no es corriente, no es claramente irracional. El éxtasis no dura, sino que declina y decae. Si una pareja está en éxtasis, pueden considerar con toda razón que su declive natural es algo muy indeseable, algo que muy bien merece ser evitado. Al cortar en seco su éxtasis, la muerte garantizaría que no va a decaer. Para una pareja así, puede que la muerte merezca ser deseada.

Hay otros modos en que deseos en apariencia locos puede que no sean irracionales. El objeto de estos deseos, por ejemplo, puede ser estéticamente atractivo. Consideremos los *caprichos*. Nagel escribe: «Uno podría, sin ninguna razón en absoluto, concebir el deseo de que hubiera perejil en la luna, y hacer todo lo que pudiera para pasar de contrabando cierta cantidad de perejil en el próximo cohete disponible; a uno simplemente podría gustarle la idea» [5]. Este deseo no es irracional. Es un capricho excelente. (Que haya perejil en el mar, por el contrario, es un capricho pobre.)

Es irracional desear algo que en ningún aspecto merece ser deseado, o merece ser evitado. Aunque podemos imaginar tales deseos con facilidad, puede que haya pocos deseos reales que sean irracionales de este modo. Y hay una gran clase de deseos que no pueden

[5] Nagel (1), p. 45.

ser irracionales. Estos son los deseos que están implicados en los dolores y placeres puramente físicos. Yo amo las duchas frías, otros las odian. En ninguno de los dos casos se es irracional. Si quiero comer algo por lo bien que me sabe, este deseo no puede ser irracional. No es irracional aunque lo que a mí me gusta disguste a todos los demás. Consideremos a continuación las experiencias que encontramos desagradables. Mucha gente tiene el intenso deseo de no escuchar el sonido que produce una tiza chirriante. Este deseo es extraño, puesto que a estas personas no les importa escuchar otros chirridos que son muy similares en timbre y tono. Pero este deseo no es irracional. Afirmaciones similares valen de lo que encontramos doloroso.

Volvamos ahora a nuestros deseos acerca de diferentes dolores y placeres posibles. Es a este nivel secundario que la acusación de «irracional» puede ser plausible en mayor medida. Alguien no es irracional simplemente por encontrar una experiencia más dolorosa que otra. Pero puede ser irracional si, cuando tiene que sufrir una de esas dos experiencias, prefiere la que será más dolorosa. Esta persona puede ser capaz de defender su preferencia. Puede creer que debe sufrir el peor dolor como una forma de penitencia. O puede querer hacerse más fuerte, más capaz de soportar futuros dolores. O puede creer que al elegir deliberadamente sufrir ahora el peor de dos dolores, y mantenerse fiel a esta elección, va a fortalecer el poder de su voluntad. O puede creer que el sufrimiento mayor le reportará sabiduría. De estos y de otros modos, el deseo que tenga alguien de sufrir el peor de dos dolores puede que no sea irracional.

Consideremos a continuación este caso imaginario. Determinado hedonista está enormemente interesado en la cualidad de sus experiencias futuras. Con una excepción, se preocupa igualmente por todas las partes de su futuro. La excepción es que él tiene *Indiferencia-hacia-el-Futuro-Martes*. En cada martes se cuida de la manera normal de lo que le ocurre. Pero nunca se preocupa por dolores o placeres posibles en un martes *futuro*. De modo que elegiría una operación dolorosa el próximo martes en vez de una operación

mucho menos dolorosa el próximo miércoles. Esta elección no sería resultado de creencia falsa alguna. Este hombre sabe que la operación será mucho más dolorosa si es el martes. Tampoco tiene falsas creencias sobre la identidad personal. Está de acuerdo en que será evidentemente el mismo que es ahora el que va a sufrir el martes. Tampoco tiene falsas creencias acerca del tiempo. Sabe que el martes es meramente una parte de un calendario convencional, con un nombre arbitrario tomado de una religión falsa. Tampoco tiene ninguna otra creencia que pudiera servirle para justificar su indiferencia al dolor en los futuros martes. Esta indiferencia es un hecho desnudo. Cuando planea su futuro, simplemente ocurre que siempre prefiere la perspectiva de un gran sufrimiento un martes que el más suave de los dolores otro día de la semana.

El patrón de intereses de este hombre *es* irracional. ¿Por qué prefiere un dolor horroroso el martes a un ligero dolor cualquier otro día? Simplemente porque el dolor horroroso será el martes. *Esto no es una razón.* Si alguien tiene que elegir entre sufrir un dolor horroroso el martes o sufrir un dolor ligero el miércoles, el hecho de que el dolor horroroso sea el martes no es razón para preferirlo. Preferir el *peor* de dos dolores por *ninguna* razón es irracional.

Puede objetarse que, puesto que la preferencia de este hombre es puramente imaginaria y es tan extraña, no podemos discutir con utilidad si es irracional. Por eso compararé otras dos actitudes ante el tiempo. Una es extremadamente común: preocuparse más por el futuro más próximo. Llamémosla *predisposición hacia lo próximo*. Aquel que tenga esta predisposición puede elegir a sabiendas tener un dolor peor unas pocas semanas más tarde en vez de un dolor menor esta tarde. Este tipo de elección se hace a menudo. Si el peor de dos dolores fuera más distante en el futuro, ¿podría ser ésta una razón para elegir ese dolor? ¿Es irracional la predisposición hacia lo próximo? Muchos autores afirman que lo es.

Consideremos ahora a alguien que tenga una *predisposición hacia el año siguiente*. Este hombre se preocupa igualmente por su futuro a lo largo del año siguiente, y se preocupa la mitad por el resto de su futuro. Una vez más, este hombre que imaginamos no tiene creencias falsas acerca del tiempo o de la identidad personal, o de ninguna otra cosa. Sabe que será justamente el mismo de ahora el que estará vivo más de un año después, y que los dolores en los años posteriores serán exactamente igual de dolorosos.

Nadie tiene el patrón de intereses de este hombre. Pero se parece mucho al patrón que es común: la predisposición hacia lo próximo. La diferencia está en que esta predisposición común es proporcional al rasgo que favorece. Los que tienen esta predisposición se preocupan más por lo que está en el futuro más cercano. Mi hombre imaginario tiene la predisposición hacia lo próximo no en una forma proporcional sino en una forma de dos pasos que es más tosca. La predisposición de este hombre traza una línea arbitraria. Se preocupa en la misma medida por los próximos 12 meses, y la mitad de eso por cualquier mes posterior. Así que elegiría a sabiendas 3 semanas de dolor dentro de 13 meses antes que 2 semanas de dolor dentro de 11 meses. Preguntado por la razón por la que prefiere el suplicio más largo, responde, «Porque queda más de un año para que llegue». Esto es como la afirmación, «Porque está más distante en el futuro». Pero está más expuesto a la crítica racional. Si un dolor va a estar más distante en el futuro, es quizá defendible pensar que esto es una razón para preocuparse menos de este dolor en este momento. Pero es difícil creer que pueda ser racional tanto preocuparse lo mismo por todos los dolores que estén dentro de los próximos 12 meses, como preocuparse sólo la mitad de eso por todos los dolores posteriores. Si un dolor va a ser sentido 53 semanas más tarde en vez de 52 semanas más tarde, ¿cómo puede esto ser una razón para preocuparse por él sólo la mitad?

Dos casos similares serían estos. Mucha gente se preocupa más por lo que les ocurre a sus vecinos o a los miembros de su propia comunidad. Muy pocos dirían que este patrón de intereses es irracional. Hay quien afirma que está moralmente requerido. Pero consideremos un hombre cuya patrón de intereses es el *Altruismo-en-un-Radio-de-una-Milla*. Este hombre se preocupa muchísimo por el bienestar de todos aquellos que están a menos de una milla de su casa, pero poco por los que están más lejos. Si se entera de que un incendio o una inundación ha afectado a la gente en un radio de una milla, donará generosamente su dinero a un fondo para ayudar a esa

gente. Pero no la ayudará si se halla a una milla y un cuarto de distancia. Esto no es una política elegida para imponer un límite a la caridad de este hombre. Es el resultado de una diferencia real en cuánto se preocupa este hombre por el sufrimiento de los demás.

El patrón de intereses de este hombre se parece a grandes rasgos al patrón que es común: una preocupación mayor por los miembros de nuestra propia comunidad. Pero su preocupación traza otra línea arbitraria. Si a alguien *no* le preocupan en absoluto los demás, esto, aunque deplorable, puede no ser irracional. Si alguien está preocupado en la misma medida por lo que les ocurre a todos, o está más preocupado por lo que les ocurre a los miembros de su propia comunidad, ninguna de estas cosas es irracional. Pero si alguien está muy preocupado por lo que les ocurre a aquellos que están a menos de una milla, y mucho menos preocupado por aquellos que viven más lejos, este patrón de intereses es irracional. ¿Cómo podría suponer una diferencia el que uno de dos desconocidos que sufren se encuentra dentro de la milla y otro más allá de la milla? El que uno de los dos se encuentre a más de una milla de distancia no es razón para estar menos preocupado.

Los seguidores de Hume afirman que, si un deseo o patrón de intereses no implica irracionalidad teórica, no puede exponerse a crítica racional. He negado esta tesis. Es verdadera cuando la aplicábamos a los deseos que están involucrados en los dolores y los placeres físicos. Pero puede no serlo de algunos deseos de primer orden. Algunos de estos pueden ser irracionales. Un ejemplo puede ser el deseo de saltar cuando se está al borde de un precipicio. Si este no es el deseo de remontar el vuelo, o de evitar el declive del éxtasis, puede ser irracional.

Los mejores ejemplos los podemos encontrar cuando nos volvemos a nuestros deseos de segundo orden acerca de dolores y placeres posibles. Tales deseos son irracionales si discriminan entre placeres igualmente buenos, o dolores igualmente malos, de un modo *arbitrario*. Es irracional preocuparse menos por dolores futuros porque los vayamos a sentir un martes, o a más de un año del momento presente. Y es irracional preocuparse menos por el sufrimiento de otras personas porque se hallen a más de una milla de distancia. En estos casos el interés no es menor a causa de una diferencia intrínseca en el objeto de interés. El interés es menor a causa de una propiedad que es puramente relativa a la posición, y que traza una línea arbitraria. Estos son los patrones de interés que son irracionales del modo más claro. Estos patrones de interés son imaginarios. Pero son versiones más toscas de patrones que son muy comunes. Mucha gente se preocupa menos por dolores futuros si están más lejos en el futuro. Y a menudo se afirma que esto es irracional. Discutiré esta afirmación en el capítulo 8 [6].

47. Tres teorías en competencia

Volvamos ahora a la teoría del Propio Interés. ¿Cómo cuestiona a PI la teoría del fin Presente, P? PI sostiene tanto

> (PI2) Lo que cada uno de nosotros tiene más razón para hacer es lo que va a ser mejor para sí mismo.

Como

> (PI3) Es irracional para todos hacer lo que uno cree que va a ser peor para sí mismo.

Un argumento a favor de P puede forzar a PI a retirarse a afirmaciones más débiles. La gravedad de las amenazas a PI depende así de dos cosas: cómo son de poderosos los argumentos, y cuán lejos, si tienen éxito, obligarán a retirarse a PI.

La amenaza más ambiciosa sería un argumento que mostrara que, en el momento en que PI entra en conflicto con P, no tenemos

[6] Los «Axiomas Evidentes en sí mismos» de Sidgwick conllevan todos, como observa Schneewind, la supresión de limitaciones *arbitrarias* al alcance o la fuerza de las razones para actuar (Schneewind, p. 300). Aunque el cargo de «irracional» se justifica *solamente* cuando un patrón de intereses traza una línea arbitraria, esto tendría amplias implicaciones.

razón para seguir PI. No tenemos razón para actuar en nuestro propio interés si esto va a frustrar lo que, en el momento de actuar, conociendo los hechos y pensando con claridad, nosotros queremos o valoramos en mayor medida. Esto sería, para PI, una derrota total.

Creo que mis argumentos justifican una versión de esta conclusión. Pero pueden mostrar algo menor. Pueden mostrar sólo que, cuando PI y P entran en conflicto, sería racional seguir cualquiera de las dos. Aunque esta es una conclusión más débil, afirmaré que, para PI, es casi igual de dañina.

Avanzaré varios argumentos. Éstos pueden ser presentados con una metáfora estratégica. Como veremos, la teoría del Propio Interés se coloca entre la moralidad y la teoría del fin Presente. Por tanto hace frente a un peligro clásico: la guerra en dos frentes. Mientras que tal vez pudiera sobrevivir al ataque desde una sola dirección, puede ser incapaz de sobrevivir a un ataque doble. Creo que esto es lo que ocurre. Muchos autores defienden que la moralidad proporciona las razones mejores o más fuertes para actuar. Al rechazar estos argumentos, un teórico del Propio Interés hace asunciones que pueden ser vueltas contra él por un teórico del fin Presente. Y sus respuestas al teórico del fin Presente, si es que son válidas, socavan su rechazo de la moralidad.

Digamos que, según nuestra manera de ver, una teoría *sobrevive* si pensamos que es racional actuar de acuerdo con ella. Y una teoría *gana* si es la única superviviente. Entonces pensaremos que es irracional *no* actuar de acuerdo con esta teoría. Si una teoría no gana, teniendo que reconocer a rivales no derrotados, tiene que cualificar sus afirmaciones. Con tres teorías, podría haber ocho resultados. Las supervivientes podrían ser:

(1)	Moralidad	La teoría del Propio Interés	La teoría del fin Presente	Triarquía
(2)		La teoría del Propio Interés	La teoría del fin Presente	Diarquías
(3)	Moralidad		La teoría del fin Presente	
(4)	Moralidad	La teoría del Propio Interés		
(5)	Moralidad			Monarquías
(6)		La teoría del Propio Interés		
(7)			La teoría del fin Presente	
(8)				Anarquía

Según la más débil de las conclusiones expuestas arriba, la teoría del Propio Interés no puede derrotar a la teoría del fin Presente. Si PI sobrevive, también P. Esto elimina (4) y (6). PI sobrevive sólo en (1) y (2). Mi conclusión más fuerte las elimina. Y afirmaré que, si PI sobrevive sólo en (1) y (2), esto significa una derrota para PI. [Este libro dice poco sobre (3), (5), (7) o el más desolado (8).]

Para alcanzar mi primer argumento, tenemos que evitar ciertos errores. Estos son más difíciles de evitar si, como muchos autores, olvidamos que la teoría del Propio Interés tiene dos rivales —que la desafían tanto las teorías morales como la teoría del fin Presente—. Si comparamos la teoría del Propio Interés con sólo una de sus rivales, podemos no darnos cuenta de cuándo roba argumentos de la otra.

48. El egoísmo psicológico

Un error es asumir que las teorías del Propio Interés y del fin Presente siempre coinciden. Nadie asume esto en el caso de la versión Instrumental de P. Lo que la gente en realidad quiere va demasiado a menudo flagrantemente en contra de sus intereses. Pero se asume por mucha gente que lo que cada persona querría más, si realmente conociera los hechos y pensara con claridad, sería hacer lo que fuera mejor para ella, o lo que mejor promoviera su propio interés a largo plazo. Esta asunción se llama *Egoísmo Psicológico*. Si

hacemos esta asunción, puede ser natural considerar P como una simple parte de PI. Aunque natural, esto sería otro error. Aunque siempre coincidieran, las dos teorías seguirían siendo distintas. Y, si sumergimos P en PI, podemos fallar a la hora de juzgar PI según sus propios méritos. Parte de su plausibilidad la puede robar de P.

Se puede hacer al Egoísmo Psicológico verdadero por definición. Algunos autores afirman que, si alguien quiere hacer lo que sabe será peor para él, no puede estar pensando «con claridad», y tiene que estar sujeto a alguna «influencia distorsionante». Cuando se hace verdadera a la afirmación de este modo, se convierte en trivial. P pierde su independencia, y por definición viene a coincidir con PI. No vale la pena discutir esta versión de P.

Hay otros dos modos en que el Egoísmo Psicológico se ha hecho verdadero por definición. Algunos autores afirman (1) que lo que será mejor para cada uno es por definición lo que en ese momento, conociendo los hechos y pensando con claridad, esta persona quiere más. Otros autores afirman (2) que, si determinado acto realizara de la mejor manera los deseos presentes de alguien, este acto, por definición, maximizaría la utilidad de esta persona. Una vez más, cuando al Egoísmo Psicológico se lo hace verdadero por definición, se convierte en trivial. Según estas dos definiciones es PI la que pierde su independencia. (1) hace coincidir a PI con la versión Deliberativa de P; (2) hace coincidir a PI con la versión Instrumental. Estas dos versiones de PI no vale la pena que las discutamos. Está claro que la definición (2) no trata sobre lo que está en nuestro propio interés a largo plazo, según cualquier teoría plausible del propio interés. Como argumentaré ahora, lo mismo es verdadero de la definición (1) [7].

La mayoría de nosotros, la mayor parte del tiempo, queremos intensamente actuar a favor de nuestros propios intereses. Pero hay muchos casos en que éste no es el deseo más intenso de uno, o en que, aunque lo sea, está contrarrestado por otros diversos deseos. Hay muchos casos en que esto es cierto incluso de alguien que conoce los hechos relevantes y piensa con claridad. Esto es así, por

[7] Véase también Gosling, Sen (2) y (4), y Broome (2).

ejemplo, cuando la teoría del fin Presente apoya a la moralidad en un conflicto con la teoría del Propio Interés. Lo que alguien quiere en la mayor medida puede ser cumplir con su deber, aunque sepa que irá en contra de sus intereses. (Recordemos que, por simplicidad, estamos considerando casos en que lo que uno quiere en la mayor medida, una vez consideradas todas las cosas, es lo mismo que lo que realizaría en la mayor medida sus deseos presentes.) Hay muchos otros casos, que no implican a la moralidad, en que lo que uno quiere en la mayor medida no coincidiría, incluso después de una deliberación ideal, con lo que de la manera más efectiva promovería el propio interés a largo plazo de esta persona. Muchos de estos casos son discutibles, u oscuros. Pero, como veremos, hay muchos otros que son claros.

Hasta qué punto son comunes los casos depende en parte de nuestra teoría del propio interés. Como afirmo en el Apéndice C, estos casos son más comunes si adoptamos la Teoría Hedonista, menos comunes si adoptamos la Teoría del Éxito. Los casos pueden ser más raros si adoptamos la Teoría No Restringida de la Realización de Deseos. Según esta teoría, la realización de cualquiera de mis deseos cuenta directamente como yendo a favor de mis intereses. Lo que va más a favor de mis intereses es lo que en la mayor medida realizaría, o me permitiría realizar, todos mis deseos a lo largo de mi vida entera. ¿Será esto siempre lo mismo que lo que realizaría mejor mis deseos presentes, si conociera la verdad y pensara con claridad? Puede haber algunos para los que estos dos siempre coincidirían. Pero hay muchos otros para los que estos dos a menudo entran en conflicto. En la vida de esas personas, PI a menudo entra en conflicto con P, aunque PI asuma la Teoría de la Realización de Deseos No Restringida. PI y P entran en conflicto porque los deseos más intensos de estas personas no son los mismos a lo largo de sus vidas.

Hay una complicación. Esta afecta a la gente que discutí en la Sección 3: esos para quienes PI es indirectamente contraproducente. En sus afirmaciones acerca de esta gente, PI entra en conflicto con P de un modo menos directo. Pero podemos ignorar esta complicación. Podemos discutir PI en casos en que no es indirectamen-

te contraproducente. Puede ser injusto para PI que nos concentremos en estos casos. Y las cuestiones importantes toman aquí una forma más clara.

El Egoísmo Psicológico no puede sobrevivir a una discusión cuidadosa. Según todas las teorías plausibles acerca del propio interés, PI y P entran en conflicto a menudo. Lo que realizaría mejor nuestros diversos deseos, en el momento de actuar, a menudo no logra coincidir con lo que de la manera más efectiva promovería nuestro propio interés a largo plazo.

49. La teoría del Propio Interés y la Moralidad

PI y P están relacionadas de una manera simple. Ambas son teorías de la racionalidad. PI se halla en una relación más sutil con la moralidad. Una teoría moral no pregunta «¿Qué es racional?», sino «¿Qué es correcto?». Sidgwick pensó que estas dos cuestiones eran, al final, la misma, puesto que las dos trataban de lo que teníamos más razón para hacer. Por eso llamó al Egoísmo uno de los «Métodos de la Ética». Un siglo después, estas dos cuestiones parecen muy separadas. Hemos expulsado al Egoísmo de la Ética, y ahora dudamos que actuar moralmente esté «requerido por la Razón». La moralidad y la teoría del Propio Interés todavía están en conflicto. Hay muchos casos en que sería mejor para uno obrar mal. En tales casos tenemos que decidir qué hacer. Tenemos que elegir entre la moralidad y PI. Pero esta elección les ha parecido a algunos imposible de discutir. Las afirmaciones de cada rival han parecido desconectadas de las afirmaciones del otro.

Pero pueden reunirse. Entre las razones para actuar, incluimos tanto razones morales como razones interesadas. Por eso podemos preguntar cuál de estas dos clases de razones es la más fuerte, o tiene más peso. Como he afirmado, podemos sospechar que esta pregunta no tiene respuesta. Podemos sospechar que no hay una escala neutral con la que poder pesar estas dos clases de razones. Pero no despachemos la cuestión como sin sentido. Podríamos

alcanzar una respuesta *sin* encontrar una escala neutral. Podemos encontrar argumentos que puedan derrotar a la teoría del Propio Interés, mostrando que sus razones carecen de peso. En la Primera Parte discutí un argumento de estos, la afirmación de que PI es contraproducente. Este argumento falló. Pero presentaré otros argumentos, y creo que al menos uno de ellos triunfa.

A estos argumentos les ayudará una explicación de la fuerza o el peso de las razones morales. Por consiguiente deberíamos incluir dentro de nuestra teoría moral una explicación de la racionalidad, y de las razones para actuar. Como esta parte de nuestra teoría moral estará interesada en lo que es racional antes que en lo que es correcto, necesita extenderse con más amplitud que el resto de nuestra teoría. En particular, necesita cometer lo que puede parecer un error. Necesita llevar bajo su extensión razones para actuar que en sí mismas no son razones morales.

Esto es lo que vienen haciendo de la manera más obvia las teorías que son *neutrales respecto del agente*, las que dan a todos los agentes sólo fines comunes. Cuando discuten la moralidad, los Neutralistas pueden tratar la teoría del Propio Interés de una manera convencional. Pueden considerarla como una teoría independiente o no moral, que tiene que ser rechazada cuando entre en conflicto con la moralidad, pero que tiene su propia esfera de influencia: la propia vida del agente, en la medida en que no afecta a los demás. Pero, cuando discuten la racionalidad, los Neutralistas se anexionan la teoría del Propio Interés, llamándola usualmente *Prudencia*. Se convierte nada más que en un caso especial derivado. La Prudencia es la rama local de la Benevolencia Racional. Esto lo afirman no sólo algunos utilitaristas, sino también ciertos no utilitaristas, como Nagel [8].

Ahora puedo describir otro error. Puede que los Neutralistas estén equivocados al anexionarse PI, pero al menos han visto lo que algunos moralistas ignoran. Han visto que la moral y las razones del propio interés pueden tener rasgos comunes, o raíces comunes. Esto es más probable en los casos en que estas razones no entran en

[8] Nagel (1).

conflicto. Por eso tales casos pueden ser engañosos. En ellos, PI puede parecer más plausible de lo que realmente es.

El caso más engañoso es aquel en que los actos de una persona le afectarán sólo a sí misma. Muchos de nosotros pensamos que, en tales casos, la moralidad no se pronuncia. Si no hay cuestión moral aquí en lo que esta persona hace, la moralidad ni entra ni deja de entrar en conflicto con PI. Pero, en sus explicaciones de la *racionalidad*, las dos pueden coincidir aquí. Si esta persona sigue PI, al hacer lo que será mejor para ella también estará haciendo lo que será mejor para todo el que este interesado. Esto es trivialmente verdadero, puesto que ella es la única que está interesada. Pero esta verdad no es en sí misma trivial. Puede llevarnos a la conclusión de que esta persona está haciendo lo que, imparcialmente considerado, produce el mejor resultado. PI puede entonces aparecer con traje prestado. Un teórico del Propio Interés puede afirmar que sería irracional para esta persona obrar de otro modo, porque con ello produciría un resultado peor. Pero el teórico de PI no tiene derecho a hacer esta afirmación. De acuerdo con PI, es con frecuencia racional para uno producir un resultado peor. Esto ocurre cuando lo que produce un resultado peor también cambia los malos efectos sobre alguien más.

Aunque no nos dejemos engañar por casos como estos, no puede haber objeción a dejarlos de lado. Podemos discutir PI en los casos en que entra en conflicto con la moralidad. Una vez más, esta precaución no puede ser injusta con PI.

50. Mi primer argumento

Antes de empezar a criticar PI, plantearé una cuestion general. Puede que algunas afirmaciones mías parezcan inverosímiles. Esto es lo que deberíamos esperar, aunque sean correctas. La teoría del Propio Interés ha dominado desde hace mucho tiempo. Se ha dado por descontado, durante más de dos milenios, que es irracional para cualquiera hacer lo que sabe que será peor para sí mismo. Los cristianos lo han dado por descontado puesto que, si el Cristianismo es

verdadero, la moralidad y el propio interés coinciden. Si los malhechores saben que irán al Infierno, cada uno de ellos sabrá que, obrando mal, está haciendo lo que va a ser peor para sí mismo. Los cristianos se han complacido en apelar a la teoría del Propio Interés, puesto que según sus asunciones PI implica que los bellacos son tontos. Observaciones similares se aplican a los musulmanes, a muchos budistas y a los hindúes. Como PI se ha enseñado durante más de dos milenios, tenemos que esperar encontrar algún eco en nuestras intuiciones. PI no puede quedar justificada simplemente por una apelación a las intuiciones que su enseñanza puede haber producido.

Como se dijo en mis dos últimas secciones, PI puede entrar en conflicto tanto con la moralidad como con la teoría del fin Presente. Estos son los casos en que PI puede ser juzgada mejor. Mi primer argumento emerge naturalmente de los rasgos definitorios de estos casos.

Cuando PI entra en conflicto con la moralidad, PI le dice a cada uno de nosotros que dé un peso supremo a sus propios intereses. Cada uno tiene que estar gobernado por el deseo de que su vida vaya, para él, lo mejor posible. Cada uno tiene que ser gobernado por este deseo, *cualquiera que sean* los costes para los demás. Por eso llamaré a este deseo *inclinación en favor de uno mismo.*

La mayoría de nosotros tiene esta inclinación. Y a menudo es más fuerte que todos nuestros otros deseos juntos. En casos tales P apoya a PI. Pero ahora estamos suponiendo que estas dos entran en conflicto. Estamos considerando a personas que, aunque conocen los hechos y piensan con claridad, *no quieren* dar un peso supremo a su propio interés. Están interesados en sus propios intereses. Pero esto o bien no es su deseo más fuerte, o bien, si lo es, está contrarrestado por sus otros deseos. De una de estas maneras, para estas personas, P falla en coincidir con PI. Lo que mejor realizaría sus deseos presentes no es lo mismo que lo que mejor promovería su propio interés a largo plazo.

PI afirma que estas personas siempre deberían ser gobernadas por la inclinación a favor de uno mismo. Deberían ser gobernadas por este deseo aunque este *no* sea su deseo más fuerte. (Esto es lo

que PI afirma en los casos más simples donde no es indirectamente contraproducente.) ¿Deberíamos aceptar esta afirmación?

Será de ayuda reformular esta cuestión. Hay diferentes versiones de la teoría Crítica del fin Presente. Una versión *coincide con PI*. ¿Esta es la versión que deberíamos aceptar? Al responder a esta pregunta veremos con más claridad lo que está implicado en aceptar PI.

De acuerdo con CP, ciertos deseos pueden ser requeridos racionalmente. Si un deseo es requerido racionalmente, cada uno de nosotros tiene una razón para hacer que este deseo se realice. Tenemos esta razón aunque *no* tengamos este deseo. Si hay un deseo que se requiere que sea nuestro *fin último*, lo que tenemos más razón para hacer es lo que quiera que cause que este deseo se realice mejor.

Para hacer que CP coincida con PI, tenemos que afirmar

> CPPI: Cada uno de nosotros está racionalmente requerido para preocuparse de su propio interés, y este deseo es supremamente racional. Es irracional preocuparse en la misma medida de cualquier otra cosa.

Según esta versión de CP, lo que cada uno de nosotros tiene más razón para hacer es todo lo que promueva mejor su propio interés.

Ahora puedo formular mi *Primer Argumento*. Deberíamos rechazar CPPI. La inclinación a favor de uno mismo *no* es supremamente racional. Deberíamos aceptar

> (CP2) Hay al menos un deseo que no es irracional, y no es menos racional que la inclinación a favor de uno mismo. Es el deseo de hacer lo que va a favor de los intereses de otras personas, cuando esto es o bien moralmente admirable, o bien el deber moral de uno.

Esta versión de CP entra en conflicto con PI. Consideremos

> *Mi Heroica Muerte*. Elijo morir de un modo que sé que será doloroso pero que salvará las vidas de otras varias personas. Estoy haciendo aquello que, conociendo los hechos y pensando con claridad, más quiero hacer, y aquello que mejor va a realizar mis deseos presentes. (En todos mis ejemplos estos dos coinciden.) También sé que estoy haciendo lo que será peor para mí. Si no sacrificara mi vida para salvar a estas personas, no me obsesionaría el remordimiento. Lo que me quedase de vida sería perfectamente digno de ser vivido.

Según esta versión de CP, mi acto es racional. Sacrifico mi vida porque, aunque me preocupa mi propia supervivencia, todavía me preocupa más la supervivencia de esas otras personas. De acuerdo con (CP2), este deseo no es menos racional que la inclinación a favor de uno mismo. De acuerdo con CP, dados los demás detalles del caso, es racional para mí realizar este deseo. Es por tanto racional para mí hacer lo que sé va a ser peor para mí.

51. LA PRIMERA RESPUESTA DEL TEÓRICO PI

En el caso recién descrito, un teórico del Propio Interés tiene que afirmar que mi acto es irracional. Por tanto tiene que rechazar (CP2). Tiene que afirmar que mi deseo es menos racional que la inclinación a favor de uno mismo.

El teórico PI podría objetar: «PI es una teoría de la racionalidad, pero de la racionalidad no de *deseos*, sino de *actos*. No necesito afirmar que, en tu ejemplo, tu deseo es menos racional que la inclinación a favor de uno mismo. Sólo necesito afirmar que, dado que estás haciendo lo que sabes que será peor para ti, tu acto es irracional».

Esta es una respuesta débil. Si el teórico PI no afirma que mi deseo es menos racional, ¿por qué deberíamos aceptar su afirmación acerca de mi acto? Consideremos

> (CP3) Si hay algún deseo que (1) no sea irracional, y (2) no sea menos racional que la inclinación a favor de uno mismo, y (3) sea verdadero de alguien que, conociendo los hechos y pensando con claridad, lo que esta persona quiere en mayor medida, una vez consideradas todas las cosas, es realizar este deseo, entonces (4) sería racional para esta persona realizar este deseo.

Esta afirmación no puede ser negada convincentemente. Aunque aceptemos PI, no tenemos ninguna razón para negar (CP3), pues-

to que esta afirmación es compatible con PI. PI es la mejor teoría si la inclinación a favor de uno mismo es supremamente racional. Entonces *no* habría (aparte de esta inclinación) deseos del tipo descrito en (CP3).

El teórico PI no puede negar (CP3) convincentemente. Si no tiene otra respuesta a mi Primer Argumento, tiene que hacer afirmaciones sobre la racionalidad de deseos diferentes. Tiene que afirmar que *no hay* deseos del tipo descrito en (CP3). Tiene que apelar a CPPI, la afirmación de que la inclinación a favor de uno mismo es supremamente racional y por eso está racionalmente requerido que sea nuestro fin último. Esta apelación a CPPI la llamaré la ***Primera Respuesta del Teórico PI***. Esta respuesta contradice de forma directa mi Primer Argumento. Ahora ampliaré este argumento. Deberíamos aceptar

> (CP4) No hay sólo uno, sino varios deseos que o bien no son irracionales o bien al menos no son menos racionales que la inclinación a favor de uno mismo.

Consideremos lo que llamaré *deseos de éxito*. Estos son deseos de tener éxito al hacer lo que, en nuestro trabajo o en nuestro tiempo libre más activo, estamos tratando de hacer. Ciertos deseos de éxito pueden ser irracionales. Esto puede ser verdadero, por ejemplo, del deseo de permanecer metido en una cueva más tiempo que nadie, o el deseo de lograr notoriedad asesinando a gente. Pero consideremos a los artistas, compositores, arquitectos, escritores, o creadores de cualquier otro tipo. Estas personas pueden querer intensamente que sus creaciones sean las mejores posibles. El más intenso de sus deseos puede ser producir una obra maestra, en pintura, música, piedra o en palabras. Y los científicos o los filósofos pueden querer intensamente hacer algún descubrimiento o algún avance intelectual fundamental. Estos deseos no son menos racionales que la inclinación a favor de uno mismo. Creo que esto ocurre con muchos otros deseos. Si creemos esto, nuestra versión de CP entra en conflicto de una manera más marcada con PI.

Vale la pena subrayar que, incluso si hay varios deseos que no son irracionales, puede haber un deseo que sea supremamente racional. Esto no puede afirmarse verosímilmente de la inclinación a favor de uno mismo. Pero sí que podríamos afirmar

> CPM: Cada uno de nosotros está racionalmente requerido a cuidarse de la moralidad, y este deseo es supremamente racional. Es irracional cuidarse en la misma medida de cualquier otra cosa.

Según esta versión de CP, siempre sería irracional obrar de un modo que creyéramos moralmente incorrecto.

CPPI hace a CP coincidir con PI. La tesis similar, CPM, puede que no haga a CP coincidir con la moralidad. La diferencia es esta. PI se propone como una teoría completa de las razones para actuar, una teoría que incluye todos los casos. Algunas teorías morales hacen la misma propuesta. Este es el caso, por ejemplo, de las teorías consecuencialistas. Pero, según las teorías morales que la mayoría de nosotros aceptamos, la moralidad no proporciona las únicas razones para obrar. Según estas teorías, hay muchos casos en que podríamos obrar de varias formas diferentes, y ninguno de estos actos sería moralmente mejor que otro. En estos casos, aunque aceptemos CPM, lo que tenemos más razones para hacer dependerá en parte de cuáles sean nuestros deseos presentes.

52. Por qué la neutralidad temporal no es lo que está en juego entre PI y P

De las tres versiones de P, he estado defendiendo la versión Crítica. Como muestran las dos últimas secciones, esta versión puede ser muy diferente de las versiones Instrumental y Deliberativa, PIn y PD. De los muchos autores que rechazan P, la mayoría ignora la versión Crítica. Esto es lamentable. Muchas objeciones a PIn y a PD *no* son objeciones a CP. En esta sección discuto una de estas objeciones.

Consideremos la opinión de que podemos estar inclinados hacia lo cercano de una manera racional. Podemos, de una manera racional, preocuparnos menos de algún dolor futuro, no porque sea menos seguro, sino simplemente porque está más distante en el futuro. Según esta opinión, la racionalidad no requiere una preocu-

pacion temporalmente neutra por el propio interés. Puede parecer que la teoría del fin Presente es la *versión radical* de esta opinión. P apela sólo a deseos presentes, y afirma que, para obrar de una manera racional, deberíamos hacer sólo lo que mejor vaya a realizar estos deseos. Puede parecer, por consiguiente, que P es la opinión de que, para ser racionales, deberíamos preocuparnos solamente de nuestros intereses presentes, o de nuestro bienestar en el momento presente. Llamemos a esta opinión el *Egoísmo del Presente*, o *EP*.

Diferentes versiones de EP apelan a diferentes teorías acerca del propio interés. Supongamos que aceptamos la Teoría Hedonista. Entonces EP implica

> *EHP*: Lo que cada uno de nosotros tiene más razón para hacer es lo que en mayor medida vaya a mejorar la calidad de su estado mental presente.

Supongamos que tengo un dolor. Si aguanto este dolor un minuto más, desaparecerá para siempre. Si aprieto un botón, mi dolor cesará instantáneamente, pero al cabo de unos minutos volverá y continuará durante los siguientes cincuenta años. EHP me dice que apriete el botón puesto que, poniendo punto final a mi dolor presente, mejoraría la calidad de mi estado mental presente. Es irrelevante que el coste de esta mejora sea un dolor de cincuenta años. EHP asume que es irracional no estar *máximamente* interesado en el estado mental presente de uno. Es irracional estar *más* interesado en los estados mentales de uno a lo largo de los próximos cincuenta años. Esta opinión es absurda.

Recordemos a continuación la versión Instrumental de la teoría del fin Presente. Esta afirma

> PIn: Lo que cada uno de nosotros tiene más razón para hacer es lo que vaya a realizar mejor sus deseos presentes.

Como EHP, PIn se refiere esencialmente al presente. Pero estas opiniones son muy diferentes. De acuerdo con PIn, puesto que ningún deseo es irracional, y es racional hacer lo que uno cree que va a realizar mejor sus deseos presentes, podría ser racional hacer cualquier cosa. *No* hay ninguna clase de acto que tenga que ser irracional. De acuerdo con EHP, puesto que una clase de acto está siempre requerida racionalmente, *toda* otra clase de acto es irracional. Estas dos opiniones se hallan en extremos opuestos.

Si somos hedonistas, estas observaciones muestran claramente que la teoría del fin Presente y el Egoísmo del Presente son opiniones muy diferentes. Supongamos a continuación que no somos hedonistas. Supongamos que nuestra teoría sobre el propio interés es la Teoría No Restringida de la Realización de Deseos. EP puede entonces coincidir con PIn.

Escribo «puede» porque, si no somos hedonistas, el Egoísmo del Presente no puede ser una concepción posible. Supongamos que pregunto, «¿Cuál acto iría más a favor de mis intereses ahora?». De una manera natural supondríamos que esto significaba, «De los actos que son posibles para mí ahora, ¿cuál iría más a favor de mis intereses?». Pero esta no es la cuestión planteada por un egoísta del presente. Él plantea una cuestión que no es corriente, y que tiene como mucho un sentido muy forzado. Esta es, «¿Qué es lo que va más a favor de mis intereses, pero no de los míos (en general), sino de los *míos-ahora*?». Podemos pensar que esta pregunta carece de sentido, porque de esta entidad *yo-ahora* no puede afirmarse que tenga intereses. (Yo puedo estar ahora interesado en ciertas cosas. Pero esto es irrelevante.) Si no rechazamos esta cuestión, y nuestra teoría sobre el propio interés es la Teoría No Restringida de la Realización de Deseos, podríamos afirmar que lo que va más a favor de los intereses míos-ahora es lo que vaya a realizar mejor mis deseos presentes. Esto sería entonces lo que tengo la mejor razón para hacer ahora, de acuerdo con EP. Y esto es también lo que tengo la máxima razón de hacer ahora, según la versión Instrumental de la teoría del fin Presente.

Si asumimos el Hedonismo, EP es absurdo. Si asumimos la Teoría No Restringida de la Realización de Deseos, EP puede coincidir con la versión Instrumental de P. Pero esto no representa ninguna objeción para P. Deberíamos aceptar no la versión Instrumental sino la Crítica.

CP no es por completo temporalmente neutral, dado que apela a los deseos presentes del agente. Pero podríamos añadir a CP

(CP5) Cada uno de nosotros está racionalmente requerido a cuidarse de su propio interés. Y este cuidado debería ser temporalmente neutral. Cada uno de nosotros debería estar igualmente preocupado por todas las partes de su vida. Pero, aunque todos nosotros debiéramos tener esta preocupación, no tendría por qué ser nuestra preocupación dominante.

Si añadimos esta tesis, CP coincide en parte con PI. Ambas teorías concuerdan en que deberíamos estar igualmente preocupados por todas las partes de nuestras vidas. Como ambas requieren esta preocupación temporalmente neutral, este requisito *no* es lo que distingue a PI de esta versión de CP. Si creemos que es irracional cuidarse menos de nuestro futuro lejano, esto *no* aporta ninguna razón para aceptar PI antes que esta versión de CP [8ª].

En este capítulo he planteado mi Primer Argumento, y he descrito la Primera Respuesta del teórico PI. Esta respuesta afirma que la inclinación a favor de uno mismo es supremamente racional. ¿Está justificada esta afirmación? ¿Sería menos racional preocuparse más por algo diferente, como la moralidad, o los intereses de otras personas, o diversos tipos de éxito? Si la respuesta correcta es No, mi Primer Argumento da en el blanco. Como no puedo demostrar que esta sea la respuesta correcta, este argumento no es decisivo. Pero creo que tengo razón en negar que la inclinación a favor de uno mismo sea supremamente racional. Si hay un solo deseo distinto que no sea menos racional, como afirma (CP2), deberíamos aceptar por nuestra parte una versión de CP que entra en conflicto con PI. Creo que hay varios deseos así.

Deberíamos rechazar PI aunque aceptemos la afirmación de que, en nuestra preocupación por nuestro propio interés, debiéramos ser temporalmente neutrales. CP puede suscribir esta afirmación. El desacuerdo entre PI y P *no* es un desacuerdo sobre esta afirmación. Es sobre si esta preocupación debería, si somos racionales, ser siempre nuestro único fin último.

[8ª] Para una discusión adicional, véase S. Kagan: «La teoría de la racionalidad del fin Presente», y mi respuesta en *Ethics*, julio de 1986.

En el próximo capítulo plantearé otro argumento contra PI. Este situará mi Primer Argumento en un contexto más amplio. El segundo argumento desafía la teoría del Propio Interés de una manera más sistemática.

7

LA APELACIÓN A LA RELATIVIDAD PLENA

53. La segunda respuesta del teórico PI

Un teórico del Propio Interés podría dar otra respuesta a mi Primer Argumento. Será útil asumir que este teórico PI acepta la Teoría de la Realización de Deseos acerca del propio interés. Esto simplificará su respuesta. No necesitamos asumir que el teórico PI acepta la Teoría No Restringida de la Realización de Deseos. Podría aceptar la Teoría del Éxito, que apela sólo a nuestros deseos sobre nuestras propias vidas. Según la Teoría del Éxito, es irrelevante que nuestros otros deseos se realicen. Podemos suponer que «deseos» significa «deseos relevantes». 273

El teórico PI podría rechazar lo que llamo su Primera Respuesta. Podría afirmar

> (PI7) La teoría del Propio Interés *no* tiene por qué asumir que la inclinación en favor de uno mismo sea supremamente racional. Hay una respuesta diferente a tu Primer Argumento. La fuerza de cualquier razón *se extiende temporalmente*. Tú tendrás más tarde razones para tratar de realizar tus deseos futuros. Como *tendrás* estas razones, tienes estas razones *ahora*. Por eso deberías rechazar la teoría

del fin Presente, que te dice que trates de realizar sólo tus deseos presentes. Lo que tienes la mejor razón para hacer es lo que va a realizar mejor, o va a ponerte mejor en condiciones de realizar, *todos* tus deseos a lo largo de tu vida.

Esta es la *Segunda Respuesta del teórico PI*. No saldrá victoriosa.

54. Las sugerencias de Sidgwick

Sidgwick escribe:

> «Verdaderamente, desde el punto de vista de la filosofía abstracta, no veo por qué el Principio Egoísta debería pasar incuestionado en mayor medida que el Universalista. No veo por qué el axioma de Prudencia no debería cuestionarse cuando entra en conflicto con la inclinación presente, por una razón similar a la que lleva a los Egoístas a su rechazo a admitir el axioma de la Benevolencia Racional. Si el Utilitarista tiene que responder a la pregunta, "¿Por qué debería yo sacrificar mi propia felicidad en aras de una felicidad mayor de otro?", seguramente tiene que ser admisible preguntar al Egoísta, "¿Por qué debería yo sacrificar un placer presente en aras de uno mayor en el futuro? ¿Por qué debería yo preocuparme de mis propias sensaciones futuras más que de las sensaciones de otras personas?". Sin duda que al Sentido Común le parece paradójico preguntar por una razón por la que uno debiera buscar su propia felicidad en conjunto: pero no veo cómo puede ser repudiada como absurda la demanda por parte de los que adoptan las ideas de la escuela de psicólogos empírica radical, aunque esas ideas se supone comúnmente que tienen una afinidad muy estrecha con el Hedonismo Egoísta. Demos por supuesto que el Egó es simplemente un sistema de fenómenos coherentes, que el "Yo" idéntico y permanente no es un hecho sino una ficción, como mantienen Hume y sus seguidores: ¿por qué, entonces, debería una parte de la serie de sensaciones en la cual se diluye (resuelve) el Ego preocuparse por otra parte de la misma serie, en mayor medida que por otra serie? [9].»

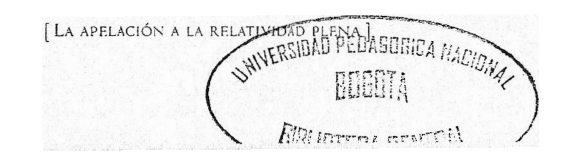
[9] Sidgwick (1), pp. 418-19.

Este pasaje tan citado carece de la claridad –la «pura luz blanca» [10] —de la mayor parte de la prosa de Sidgwick—. A mí me parece que la explicación es esta. La *Prudencia Egoísta* de Sidgwick es la teoría del Propio Interés sobre la racionalidad. Su fragmento sugiere dos argumentos contra esta teoría. El fragmento no es claro porque, al formular uno de esos argumentos, Sidgwick se equivoca.

Sidgwick primero afirma que la Prudencia y la Benevolencia Racional pueden ser cuestionadas por razones «similares». ¿Cuáles son estas razones? Las dos últimas frases sugieren una respuesta. Si son sólo «Hume y sus seguidores» los que no pueden rechazar el cuestionamiento de la Prudencia, la concepción de Hume sobre la identidad personal puede ser la «razón» de este cuestionamiento. La razón «similar» del cuestionamiento de la Benevolencia Racional puede ser alguna opinión diferente sobre la identidad personal.

Esta interpretación se corresponde con dos de las ulteriores afirmaciones de Sidgwick. «Sería contrario al Sentido Común negar que la distinción entre un individuo cualquiera y cualquier otro es real y fundamental... siendo esto así, no veo cómo puede probarse que esta distinción no tenga que ser tomada como fundamental a la hora de determinar el fin último de la conducta racional de un individuo» [11]. Estas declaraciones sugieren cómo un teórico del Propio Interés puede cuestionar el requisito de Benevolencia Racional o Imparcial que es característico del moralista. Este cuestionamiento puede apelar a la naturaleza fundamental de la distinción entre individuos, o entre diferentes vidas. La distinción entre vidas es profunda y fundamental si su correlato, la unidad de cada vida, es profundo y fundamental. Como defenderé después, esto es lo que el Sentido Común cree en lo que respecta a la identidad personal. Según esta concepción, tiene una gran significación racional y moral el que seamos personas diferentes, cada una de las cuales con su propia vida que vivir. Esto apoya la tesis de que el fin supremamente racional, para cada persona, es que su propia vida vaya lo mejor posible. Y esta es la tesis con la que un teórico del Propio

[10] Blanshard.
[11] Sidgwick (1), p. 498.



Interés puede cuestionar el requisito de la Benevolencia Imparcial característico del moralista. Como sugiere Sidgwick, este cuestionamiento vendría apoyado por la concepción de sentido común de la identidad personal.

Esta concepción es negada por «Hume y sus seguidores». Como escribe Sidgwick, Hume creía que «el Ego es meramente un sistema de fenómenos coherentes... el "Yo" idéntico y permanente no es un hecho sino una ficción». Y Sidgwick sugiere que la concepción de Hume apoya un cuestionamiento de la teoría del Propio Interés.

Los dos cuestionamientos sugeridos no pueden estar a la vez bien fundamentados. Al cuestionar el requisito de Benevolencia Imparcial, un teórico del Propio Interés apela a la concepción de sentido común de la identidad personal. La teoría del Propio Interés puede a su vez ser cuestionada por una apelación a la concepción de Hume. El primer cuestionamiento está bien fundamentado si la concepción de sentido común es verdadera. Si esta concepción es verdadera, el segundo cuestionamiento apela a una concepción falsa. Como Sidgwick aceptaba la concepción de sentido común, y creía que la concepción de Hume era falsa, no es sorprendente que no desarrollara el cuestionamiento que él mismo sugirió de la teoría del Propio Interés.

La concepción de Hume es inadecuada. Pero en la Tercera Parte defenderé una concepción que, en los aspectos relevantes, sigue a Hume. Y afirmaré que, como sugiere Sidgwick, esta concepción apoya un argumento contra la teoría del Propio Interés.

En el resto de la Segunda Parte, mi objetivo es diferente. Continuaré cuestionando la teoría del Propio Interés con argumentos que no apelan a ninguna concepción de la naturaleza de la identidad personal. Uno de estos es el segundo argumento que el fragmento de Sidgwick sugiere.

Como este fragmento emplea el ambiguo «debería», puede parecer que versa sobre la moralidad. Pero, como escribí, versa sobre lo que tenemos más razón para hacer. Podemos formular así el «axioma de la Benevolencia Racional» de Sidgwick

> *BR*: La razón requiere que cada persona aspire a la suma más grande posible de felicidad, imparcialmente considerada.

Esto puede cuestionarse, afirma Sidgwick, preguntando

> (Pr1) «¿Por qué debería yo sacrificar mi propia felicidad por la mayor felicidad de otro?».

Podemos formular así el «axioma de Prudencia» de Sidgwick

> *PIH*: La razón requiere que cada persona aspire a su propia felicidad, y a que esta sea la mayor posible.

Esto puede cuestionarse preguntando

> (Pr2) «¿Por qué debería yo sacrificar un placer presente por uno mayor en el futuro? ¿Por qué debería yo preocuparme por mis propias sensaciones futuras más que por las sensaciones de otros?».

Afirmé que estas dos preguntas pueden tener fundamentos «similares» porque cada una apela implícitamente a una concepción de la identidad personal. Esto puede ser parte de lo que Sidgwick tenía en mente, como sugieren tanto el final de su fragmento como las afirmaciones posteriores que cité. Pero hay una interpretación más simple. (Pr1) rechaza el requisito de imparcialidad entre personas diferentes. Implica que un agente racional puede dar un estatus especial a una persona particular: *él mismo*. (Pr2) rechaza el requisito de imparcialidad entre momentos temporales diferentes. Implica que un agente racional puede dar un estatus especial a un momento temporal particular: el *presente*, o el tiempo de actuar. Las dos preguntas tienen fundamentos «similares» a causa de la analogía entre uno mismo y el presente. Esta analogía proporciona otro argumento contra la teoría del Propio Interés.

55. Cómo PI es relativa de forma incompleta

Este argumento puede presentarse con estas observaciones. La teoría moral de Sidgwick requiere lo que él llama Benevolencia

Racional. Según esta teoría, un agente puede no conceder un estatus especial ni a sí mismo ni al presente. Al exigir tanto neutralidad personal como temporal, esta teoría es *pura*. Otra teoría pura es la del fin Presente, que rechaza los requisitos tanto de la neutralidad personal como de la temporal. La teoría del Propio Interés no es pura. Es una teoría *híbrida*. PI rechaza el requisito de la neutralidad personal pero requiere la neutralidad temporal. PI permite al agente darse preeminencia a sí mismo, pero insiste en que no puede dar preeminencia al momento temporal de actuar. No tiene que asignar un peso especial a lo que él quiere o valora *ahora*. Tiene que asignar el mismo peso a todas las partes de su vida, o a lo que quiere o valora en todo tiempo.

Sidgwick puede haber visto que, como híbrida, PI puede ser acusada de una especie de inconsistencia. Si el agente tiene un estatus especial, ¿por qué negar este estatus al momento de actuar? Podemos objetar a PI que es *relativa de forma incompleta*.

De acuerdo con PI, las razones de actuar pueden ser relativas al agente. Tengo una razón para hacer todo lo que vaya a ser mejor para mí. Esta es una razón para mí pero no para ti. Tú *no* tienes una razón para hacer todo lo que vaya a ser mejor para mí.

Un teórico del fin Presente puede afirmar

> (P1) Si las razones pueden ser relativas, pueden ser *completamente* relativas: pueden ser relativas al agente *en el momento de obrar*.

Esta tesis puede apelar a la analogía entre uno mismo y el presente, o aquello a lo que se refieren las palabras «yo» y «ahora». Esta analogía se mantiene sólo a un nivel formal. Los momentos particulares de tiempo no se parecen a las personas particulares. Pero la palabra «yo» refiere a una persona particular *de la misma manera en que* la palabra «ahora» refiere a un momento temporal particular. Y cuando cada uno de nosotros está decidiendo qué hacer, se está preguntando, «¿Qué debería hacer *yo ahora*?». Dada la analogía entre «yo» y «ahora», una teoría debe dar a los dos el mismo tratamiento. Por eso (P1) afirma que una razón puede tener fuerza sólo para *mí ahora*.

«Aquí» es también análogo a «yo». Cuando un consejero me cuenta cómo escapó de un medio menos adverso, yo podría protestar, «Pero ¿qué debería hacer *yo aquí*?». Si una persona pudiera estar en varios lugares a la vez, no bastaría con preguntar «¿Qué debería hacer yo ahora?». Si yo pudiera estar ahora en varios lugares distintos, esta pregunta no sería *completamente* relativa. Pero de hecho no hay necesidad de este añadido. Basta con afirmar que una razón puede tener fuerza sólo para mí ahora. No necesitamos añadir «aquí» porque, como yo estoy aquí, no puedo estar ahora también en otro sitio [12].

Un teórico del fin Presente podría hacer dos afirmaciones más contundentes. Podría decir

> (P2) Las razones para actuar *tienen que* ser completamente relativas. Deberíamos rechazar las tesis que impliquen que las razones pueden ser relativas *de forma incompleta*. Así que deberíamos rechazar la tesis de que las razones pueden ser relativas al agente pero temporalmente neutrales. Podemos llamar a tales tesis *incompatibles con la plena relatividad*.

> (P3) Consideremos cualquier par de tesis que estén relacionadas del siguiente modo: La primera tesis contiene la palabra «yo», pero no contiene la palabra «ahora». La segunda tesis es en todo como la primera, salvo que *sí* contiene la palabra «ahora». Llamemos a un par de tesis así *análogas*. Si la primera tesis entra en conflicto con la segunda, es incompatible con la relatividad plena, y por eso debería ser rechazada. Si la primera tesis *no* entra en conflicto con la segunda, es una cuestión abierta si deberíamos aceptar la primera tesis. Pero, *si* aceptamos la primera tesis, deberíamos aceptar *también* la segunda. Esto es así porque, si aceptamos la primera pero rechazamos la segunda, nuestra concepción es incompatible con la relatividad plena. Y no estamos dando, como debiéramos hacer, a «yo» y «ahora» el mismo tratamiento.

[12] En el caso imaginario en que divido mi mente, discutido en la Sección 87, mi razón *necesitaría* ser tres veces relativa. Yo tendría que preguntar, «¿Qué debería hacer *yo ahora* en *esta* mitad de mi mente?».

(P3) puede ser poco clara, y puede ser poco claro por qué un teórico del fin Presente hace esta afirmación. Pero, cuando aplico (P3), estos dos puntos pueden volverse claros.

Las tesis de la (P1) a la (P3) dan formulación a lo que llamo la *Apelación a la Relatividad Plena*. Creo que es una poderosa objeción a la teoría del Propio Interés.

56. Cómo se equivocó Sidgwick

Antes de discutir esta objeción, ayudará sugerir por qué Sidgwick fracasó a la hora de captar su fuerza.

Al comienzo del fragmento citado arriba, Sidgwick parece enterado de que la amenaza a PI viene de P. Su segunda frase comienza: «No veo por qué el axioma de Prudencia no debería cuestionarse, *cuando entra en conflicto con la inclinación presente*...». Esto sugiere la pregunta que sería hecha por un teórico del fin Presente:

> (Pr3) ¿Por qué debería yo aspirar a mi propia felicidad mayor si esto no es lo que, en el momento de actuar, yo más quiero o valoro?

Esta es una buena pregunta. Pero no es la pregunta que Sidgwick hace después. Su fragmento sugiere que conocía a medias la Apelación a la Relatividad Plena. Pero no formula plenamente esta apelación y no la discute en detalle. Sugiero que puede haberse equivocado de las maneras siguientes:

(a) Puede o haber pasado por alto la teoría del fin Presente, o haber dado por supuesto que no era una rival seria para la teoría del Propio Interés.
(b) Él era hedonista. En su forma hedonista, PI implica

> (PI8) La razón requiere que yo aspire a mi propia felicidad mayor posible.

Sidgwick aceptaba esta tesis. Y puede haber pensado que, si apelamos a la analogía entre «yo» y «ahora», los que aceptan (PI8) deberían aceptar también

> EHP: La razón requiere que yo aspire ahora a mi propia felicidad mayor posible ahora, o en el momento presente.

Esta es otra formulación del Egoísmo Hedonista del Presente. Como he dicho, esta concepción es absurda. El carácter absurdo de EHP puede haber llevado a Sidgwick a rechazar tanto la analogía entre «yo» y «ahora» como la Apelación a la Relatividad Plena.

(c) Esta Apelación nos dice que rechacemos PI y aceptemos P. Sidgwick creía que PI es plausible. Puede no haber visto que, aunque apelemos a la relatividad plena, podemos admitir que PI es plausible. PI puede formularse en una serie de afirmaciones. Y la Apelación a la Relatividad Plena nos permite aceptar algunas de esas afirmaciones. Podemos aceptar aquellas partes de PI que sean compatibles con la teoría del fin Presente. Sidgwick pensaba que PI es plausible, y puede haber pensado que EHP es absurdo. Pero lo que es plausible en PI son las partes de PI que son tanto compatibles con, como análogas a partes de P. Y el absurdo EHP es análogo a una parte de PI que deberíamos rechazar. Puesto que esto es así, como afirmaré, la Apelación a la Relatividad Plena no entra en conflicto tan frontalmente con las intuiciones de Sidgwick.

57. La apelación aplicada a un nivel formal

Los consecuencialistas rechazan la inclinación en favor de uno mismo. Rashdall pregunta, por ejemplo, por qué «una Razón imparcial o impersonal... debería atribuir más importancia al placer de un hombre que al de otro». «Es... inteligible» [13], concede él,

[13] Excepto para G. E. Moore. «¿Qué quiere decir el profesor Sidgwick con las expresiones "el fin racional último para él mismo", y "de suma importancia *para él*"? Él no pretende definirlas; y es en gran medida el uso de tales expresiones sin definir lo que causa los absurdos que se perpetran en filosofía» (Moore, p. 99). La negación por parte de Moore de que estas expresiones tengan sentido constituye su «refutación» de Sidgwick. Y añade: «No puede desearse ninguna refutación más completa y más a fondo de ninguna teoría». Pero el mismo Moore había afirmado

«que una cosa pueda ser razonablemente deseada desde el punto de vista particular de un hombre, y otra cosa diferente cuando se adopta el punto de vista de un todo más amplio. Pero, ¿pueden ser ambos puntos de vista igualmente razonables? ¿Cómo puede ser razonable adoptar el punto de vista de la parte una vez que el hombre conoce la existencia del todo...?» [14].

Un teórico del Propio Interés puede contestar: «No estamos preguntando por lo que sería racional para la *Razón* hacer. Estamos preguntando lo que es racional hacer para *mí*. Y yo puedo razonablemente rehusarme a adoptar "el punto de vista del todo". Yo no soy el todo. ¿Por qué no puede ser *mi* punto de vista, precisamente, *mi* punto de vista?».

La teoría del Propio Interés puede a su vez ser cuestionada. Un teórico del fin Presente puede decir: «No estamos preguntando sólo qué es racional hacer para mí. Estamos preguntando qué es racional hacer para mí *ahora*. Tenemos que considerar, no sólo mi punto de vista, sino mi punto de vista presente». Como escribe Williams, «La perspectiva correcta sobre la vida de uno es la que se toma *desde ahora*» [15].

Se puede presentar esta cuestión en términos más formales. Siguiendo a Nagel, distinguí dos clases de razones para actuar. Nagel llama a una razón *objetiva* si no está atada a ningún punto de vista. Supongamos que se afirma que hay una razón para aliviar el sufrimiento de cierta persona. Esta razón es objetiva si es una razón para todos —para cualquiera que pudiera aliviar el sufrimiento de esta persona—. Llamo a tales razones neutrales respecto del agente. Las razones *subjetivas* de Nagel son razones sólo para el agente. A estas las llamo relativas al agente [16].

Debiera explicar con más detalle el sentido en que las razones pueden ser relativas. En un sentido, todas las razones pueden ser relativas a un agente, y a un momento temporal y a un lugar. Aunque

poco antes que "bueno" no podía definirse. Podemos exclamar lo que exclamó Mackie con toda justicia [Mackie (1), p. 323], «¡Qué descaro!».

[14] Rashdall, p. 45.

[15] Williams (3), p. 209.

[16] Nagel (1) y (3).

tú y yo tratemos de lograr algún fin común, podemos estar en diferentes situaciones causales. Yo puedo tener una razón para actuar de un modo que promueva nuestro fin común, pero puede que tú no tengas tal razón, dado que quizá seas incapaz de obrar de ese modo. Como hasta las razones neutrales respecto del agente pueden ser, en este sentido, relativas al agente, este sentido es irrelevante para esta discusión.

Cuando digo de una razón que es relativa al agente, no estoy afirmando que esta razón *no pueda* ser una razón para otros agentes. Todo lo que estoy diciendo es que puede no serlo. Según la teoría del fin Presente, mi razón para actuar es una razón para otros agentes si ellos y yo tenemos el mismo fin. De modo similar, cuando P afirma que determinada razón es relativa al tiempo, la afirmación no es que, a medida que el tiempo pasa, esta razón está condenada a perderse. La afirmación es sólo que puede perderse. Se perderá si hay un cambio en el fin del agente.

Si todas las razones para actuar fueran neutrales con respecto al agente, esto sería fatal para la teoría del Propio Interés. Consideremos la razón de cada persona de promover sus propios intereses. Si esta fuera una razón para todo el mundo, cada persona tendría la misma razón para promover los intereses de todo el mundo. La teoría del Propio Interés sería anexionada por la Benevolencia Imparcial. Un teórico del Propio Interés tiene por tanto que afirmar que las razones para actuar pueden ser relativas al agente. Pueden ser razones para el agente sin ser razones para nadie más.

Un teórico del fin Presente estaría de acuerdo. Pero él puede añadir (P1): la tesis de que una razón puede ser relativa al agente en el momento de obrar. Puede ser una razón para él en ese momento sin ser una razón para él en otros momentos.

Esta tesis desafía la Segunda Respuesta del teórico PI, la que asume que la fuerza de cualquier razón se extiende a través del tiempo. De acuerdo con PI, como escribe Nagel, «*hay* una razón para promover aquello para lo que... *habrá* una razón» [17]. Las razones interesadas son en este sentido intemporales, o temporalmente neutrales.

[17] Nagel (1), p. 45 (la cursiva es mía).

Aunque intemporales, no son impersonales. Como dije, PI es una teoría híbrida. De acuerdo con las teorías morales neutralistas, las razones para actuar son tanto intemporales como impersonales. De acuerdo con la teoría del fin Presente, las razones son tanto relativas al momento temporal como relativas al agente: son razones para el agente en el momento de actuar. De acuerdo con la teoría del Propio Interés, las razones son relativas al agente, pero no son relativas al momento temporal. Aunque PI rechaza el requisito de impersonalidad, exige neutralidad temporal.

Como híbrida que es, PI puede ser atacada desde las dos direcciones. Y lo que PI dice contra un rival puede ser vuelto contra ella por el otro. Al rechazar el Neutralismo, un teórico del Propio Interés tiene que afirmar que una razón puede tener fuerza sólo para el agente. Pero los fundamentos de esta afirmación apoyan una tesis que va más lejos. Si una razón puede tener fuerza sólo para el agente, puede tener fuerza para el agente sólo en el momento de actuar. El teórico del Propio Interés tiene que rechazar esta tesis. Tiene que atacar la noción de una razón relativa al tiempo. Pero los argumentos que muestran que las razones tienen que ser temporalmente neutrales, refutando así la teoría del fin Presente, pueden también mostrar que las razones tienen que ser neutrales entre diferentes personas, refutando así la teoría del Propio Interés.

Una vez Nagel presentó un argumento de esta segunda clase. Si su argumento triunfa, el Neutralismo gana. Pero ahora no estoy discutiendo el argumento de Nagel, sino la apelación a la relatividad plena. Como el argumento de Nagel, esta apelación cuestiona la teoría del Propio Interés. Una aparte de esta apelación es (P1), la tesis de que, si las razones pueden ser relativas al agente, pueden ser plenamente relativas: relativas al agente en el momento de actuar.

O las razones pueden ser relativas, o no pueden. Si no pueden, como argumentó Nagel, el Neutralismo gana. Tenemos que rechazar tanto la teoría del Propio Interés, como la teoría del fin Presente, y la mayor parte de la Moralidad del Sentido Común.

Supongamos a renglón seguido que, como cree Nagel ahora, las razones pueden ser relativas. (P1) afirma correctamente que, si las razones pueden ser relativas, pueden ser relativas al agente en el

momento de actuar. Como defenderé en el próximo capítulo, podría ser verdad que yo *una vez* tuviera una razón para promocionar determinado fin, sin que sea verdad que tengo esta razón *ahora*. Y podría ser verdad que yo *tendré* una razón para promover determinado fin, sin que sea verdad que yo tengo esta razón *ahora*. Como estas dos cosas podrían ser verdaderas, no puede decirse que la fuerza de cualquier razón se extiende a través del tiempo. Esto socava la Segunda Respuesta del teórico PI.

Además de apelar a (P1), un teórico del fin Presente puede apelar a los más contundentes (P2) y (P3). Ahora mostraré que, si esta apelación está justificada, hay más razones para rechazar PI.

58. LA APELACIÓN APLICADA A OTRAS TESIS

Supongamos en primer lugar que alguien acepta tanto PI como la Teoría de la Realización de Deseos sobre el propio interés. Semejante persona podría formular PI, como se aplica a él mismo, con la tesis

> (PI9) Lo que tengo más razón para hacer es lo que vaya a realizar mejor todos mis deseos a lo largo de mi vida completa.

La tesis análoga es

> (P4) Lo que tengo más razón para hacer ahora es lo que vaya a realizar mejor todos los deseos que tengo ahora.

Como estas tesis chocan, la Apelación a la Relatividad Plena nos dice que rechacemos (PI9), que es incompatible con la relatividad plena. Al decirnos que rechacemos esta tesis, la Apelación nos dice que rechacemos una versión de PI. Y eso nos permite aceptar (P4), que formula la versión Instrumental de P.

Consideremos a continuación

> (PI10) Puedo racionalmente ignorar deseos que no son míos,

y

(P5) Puedo racionalmente ignorar ahora deseos que no son míos ahora.

Estas tesis no entran en conflicto. De acuerdo con la Apelación a la Relatividad Plena, deberíamos por consiguiente aceptar o ambas o ninguna. Como un teórico del Propio Interés tiene que aceptar (PI10), tiene que aceptar (P5). Pero (P5) es tanto una negación de PI como una afirmación parcial de P. Como antes, la Apelación cuenta contra PI y a favor de P.

Aunque un teórico del fin Presente rechazara la teoría del Propio Interés, aceptaría algunas de las afirmaciones que hace PI. Así, aceptaría el rechazo de BR, la tesis de que la razón requiere benevolencia imparcial. Y, aunque rechazara

PIH: La razón requiere que cada persona aspire a su propia felicidad mayor posible,

aceptaría en su lugar

(P6) No es irracional preocuparse más por la felicidad propia.

Si no es hedonista, añadiría

(P7) No es irracional preocuparse más por lo que le pasa a uno mismo, o por el propio interés de uno.

Esta tesis defiende lo que llamo la inclinación en favor de uno mismo. Pero, a diferencia de PI, no insiste en que un agente racional tenga que tener, y tenga que ser gobernado por, esta inclinación particular. Sólo afirma que esta inclinación no es irracional.

Supongamos que aceptamos (P7). La tesis análoga es

(P8) No es irracional preocuparse más por lo que le está pasando a uno mismo en el momento presente.

Esta tesis defiende lo que llamaré la *inclinación hacia el presente*. Esta es todavía más común que la inclinación a favor de uno mismo. Las dos inclinaciones pueden expresarse en pensamientos como

estos: «Sabía que alguien tenía que hacer este horrible trabajo, pero ojalá no tuviera que ser *yo*»; «Sabía que yo tenía que hacer este trabajo tarde o temprano, pero ojalá no tuviera que ser *ahora*». (O: «Sabía que me tenían que empastar la muela, pero ojalá que no tuviera que ser en este preciso momento».)

De acuerdo con la Apelación a la Relatividad Plena, si aceptamos (P7) deberíamos aceptar también (P8). Esto parece correcto. Las dos inclinaciones pueden ser defendidas con razones similares. Mis razones para preocuparme de lo que me ocurre a mí difieren en clase de mis razones para preocuparme de lo que les ocurre a otras personas. La relación entre yo y mis propias sensaciones es mucho más íntima que las relaciones entre yo y las sensaciones de otros. Este hecho hace a (P7) plausible. Del mismo modo, mis razones para preocuparme de lo que me está ocurriendo a mí difieren ahora en clase de mis razones para preocuparme ahora de lo que me ocurrió o me ocurrirá. La relación entre yo ahora y lo que estoy sintiendo ahora es mucho más íntima que la relación entre yo ahora y mis sensaciones en otros momentos temporales. Este hecho hace a (P8) plausible.

Como (P7) lo suscribirían tanto un teórico del Propio Interés como un teórico del fin Presente, el último puede decir: «La teoría del Propio Interés es en parte correcta. Su error es pasar de (P7) a una tesis más contundente. De acuerdo con (P7), no es irracional preocuparse más del propio interés de uno. Esto es plausible. Pero no es plausible afirmar que la preocupación por nuestro propio interés tiene que ser siempre, si somos racionales, nuestro fin último».

(P7) y (P8) no son centrales para la teoría del fin Presente. Si alguien no se preocupara más o de sí mismo o de sus sensaciones presentes, no sería tildado de irracional por un teórico del fin Presente. Estas dos tesis no están implicadas por la tesis muy diferente que *es* central para la versión crítica de P. (Esta es la tesis de que, si conocemos los hechos y pensamos con claridad, lo que tenemos más razón para hacer es lo que realizaría mejor aquellos de nuestros deseos presentes que no son irracionales.)

Aunque un teórico del fin Presente aceptara (P7), la incrustaría en una tesis más amplia. Esta podría ser

> (P9) Un patrón de intereses no es irracional meramente porque no dé peso supremo al logro del resultado mejor posible, imparcialmente considerado. No sería menos racional para *mí* preocuparme *más* de lo que me ocurre a *mí*, o a la gente a la que *yo* quiero, o del éxito de lo que *yo* estoy tratando de lograr, o de las causas con las que *yo* estoy comprometido.

Esta tesis no da un lugar especial a la inclinación a favor de uno mismo. Simplemente cita esta inclinación como un ejemplo de un interés que, aunque no sea imparcial ni neutral respecto del agente, no es menos racional. Este es el interés más simple de todos y el más evidentemente sesgado o relativo al agente. Pero hay muchísimos otros, algunos de los que (P9) cita, y de los que afirma que no son menos racionales.

Si aceptamos (P9), la Apelación a la Relatividad Plena nos dice que aceptemos

> (P10) Un patrón de intereses no es irracional simplemente porque no dé peso supremo al propio interés. No sería menos racional para mí preocuparme más *ahora* de las personas a las que quiero *ahora*, o del éxito de lo que estoy tratando de lograr *ahora*, o de las causas con las que estoy *ahora* comprometido. Y puede que no sea menos racional para mi preocupación por mi propio interés incluir un sesgo temporal relativo al presente. Por ejemplo, puede que no sea menos racional para mí preocuparme más *ahora* de lo que me está ocurriendo a mí *ahora*.

(P10) *es* una de las tesis centrales de la teoría del fin Presente. No concede un lugar especial a la inclinación hacia el presente. Esto se cita simplemente como un ejemplo de un interés que, aunque relativo al agente en el momento de actuar, *puede* que no sea menos racional.

Escribo «puede» porque hay diferentes versiones de CP. Una versión establece que, en mi preocupación por mi propio interés, yo

debería ser temporalmente neutral. Esto es consistente con (P10). Según esta versión de CP, yo también debería tener una preocupación temporalmente neutral por los intereses de las personas que quiero. Al tratar de hacer lo que vaya a ser lo mejor para estas personas, yo debería dar igual peso a todas las partes de sus vidas. Esto es consistente con la tesis de que no es menos racional para mí estar más preocupado *ahora* por los intereses de las personas a las que quiero *ahora*. Y, aunque mi preocupación por mi propio interés debiera ser temporalmente neutral, no es menos racional para mí preocuparme más *ahora* de lo que estoy tratando de lograr *ahora*, o de las causas con las que *ahora* estoy comprometido. Mi preocupación por este logro o estas causas, *no* es una preocupación por mi propio interés.

Si aceptamos (P10), rechazaremos la tesis central de la teoría del Propio Interés: la de que es irracional para cualquiera hacer lo que sabe será peor para él mismo. De acuerdo con PI, el único patrón de intereses supremamente racional es una inclinación temporalmente neutral a favor de uno mismo. De acuerdo con (P10), muchos otros patrones de interés no son menos racionales. Si esto es así, mi Primer Argumento triunfa. Si tenemos uno de estos otros patrones de interés, no sería menos racional actuar a partir de él. Esto no sería menos racional aun cuando, al obrar así, estuviéramos haciendo lo que sabemos que va a ir en contra de nuestro propio interés a largo plazo.

¿Deberíamos aceptar (P10)? Más exactamente, ¿podemos rechazar (P10) si hemos aceptado (P7), la tesis de que no es irracional estar predispuesto a favor de uno mismo? ¿Es tal inclinación la única o la supremamente racional? He mencionado algunos otros deseos e intereses de los que afirmé que no eran menos racionales. Si esta tesis está justificada, deberíamos aceptar una versión matizada de (P10).

Deberíamos matizar (P10) porque puede ser irracional tener ciertos deseos de éxito, o estar comprometido con ciertas causas. Así, puede ser irracional querer permanecer encerrado en una cueva más tiempo que nadie. Pero, como he dicho, hay muchos deseos de éxito que no son menos racionales que la inclinación a favor de uno

mismo. No es menos racional querer crear ciertos tipos de belleza, o lograr ciertos tipos de conocimiento. Y hay otros muchos ejemplos. Afirmaciones similares se aplican a las causas con las que estamos comprometidos. No era irracional en el siglo XIX estar comprometido con la adopción del esperanto como *lingua franca* del mundo —o lenguaje internacional—. Dadas las posiciones relativas del inglés y del esperanto, este compromiso puede ser irracional ahora. Pero hay muchas otras causas con las que comprometerse no es menos racional que la inclinación a favor de uno mismo. Como deberíamos aceptar una versión matizada de (P10), deberíamos aceptar una versión de CP que a menudo entra en conflicto con PI.

En este capítulo he defendido que, si una razón puede tener fuerza sólo para una persona, una razón puede tener fuerza para una persona sólo en un momento temporal. Deberíamos rechazar la tesis de que la fuerza de cualquier razón se extiende a través del tiempo. Deberíamos rechazar, por consiguiente, la Segunda Respuesta del teórico PI a mi Primer Argumento, que apela a esta tesis. Si no tiene otra respuesta, puede que el teórico PI tenga que volver a su Primera Respuesta. Quizás tenga que afirmar que la inclinación a favor de uno mismo es supremamente racional. Deberíamos rechazar esta tesis. Si el teórico PI no dispone de otra respuesta, deberíamos rechazar PI.

He defendido también que hay otras razones para rechazar PI. Estas nos las proporciona la Apelación a la Relatividad Plena. De acuerdo con esta apelación, las únicas teorías sostenibles son la moralidad y la teoría del fin Presente, porque sólo ellas dan a «yo» y a «ahora» el mismo tratamiento. (Las teorías morales neutrales respecto del agente dan claramente a «yo» y «ahora» el mismo tratamiento. Y también lo hacen, aunque de forma menos obvia, las teorías relativas al agente. Éstas exigen que dé un peso especial a los intereses de determinadas personas. Por ejemplo, yo debería dar un peso especial a los intereses de mis hijos. Puede parecer que esta afirmación da a «yo» un tratamiento que niega a «ahora». Pero no es así. Mi relación con mis hijos no puede darse en un determinado período de tiempo, y dejar de darse en otro período temporal distinto. No es posible que mis hijos puedan existir sin ser mis hijos.

Por esta razón, en la afirmación sobre mis obligaciones para con mis hijos, no necesitamos incluir la palabra «ahora». Hay otras relaciones que *pueden* darse en un momento determinado y dejar de darse en otros momentos. En sus tesis acerca de tales relaciones, una teoría moral relativa al agente sí que incluye la palabra «ahora». Por ejemplo, si soy médico tengo obligaciones especiales con los que son *ahora* mis pacientes. Y no tengo tales obligaciones con los que fueron una vez mis pacientes pero ahora lo son de algún otro médico.)

En el capítulo siguiente doy más razones para rechazar PI. Pero mi objetivo principal será discutir algunas cuestiones desconcertantes.

8

DIFERENTES ACTITUDES ANTE EL TIEMPO

La teoría del Propio Interés mantiene que, en nuestra preocupación por nuestro propio interés, debiéramos ser temporalmente neutrales. Como he dicho, un teórico del fin Presente puede suscribir también esta tesis. Ahora preguntaré si esta tesis esta justificada. Si la respuesta es Si, esto no va a representar objeción alguna contra P. Pero, si la respuesta es No, esto constituye otra objeción contra PI.

59. ¿Es irracional no dar peso alguno a los propios deseos pasados?

Consideremos en primer lugar a esos teóricos PI que aceptan la Teoría de la Realización de los Deseos en lo que hace al propio interés. Según todas las versiones de esta teoría, lo que es mejor de todo para alguien es aquello que vaya a realizar lo mejor posible sus deseos, a lo largo de su vida. Y la realización de los deseos de alguien es buena para él, y su no realización mala para él, aunque esta persona nunca sepa si estos deseos se han realizado.

Al decidir lo que realizaría mis deseos lo mejor posible, tenemos que tratar de predecir los deseos que yo tendría si mi vida discu-

rriera en las diferentes maneras en que podría discurrir. La realización de un deseo cuenta más si el deseo es muy intenso. ¿También debería contar más si yo lo hubiera tenido durante mucho tiempo? En el caso de los deseos intensos, parece plausible responder Sí; pero en el caso de los deseos débiles la respuesta es menos clara.

Según la Teoría No Restringida de la Realización de los Deseos, es bueno para mí que *cualquiera* de mis deseos se realice, y malo para mí que alguno no se realice. Otra versión es la *Teoría del Éxito*. Esta sólo da peso a mis deseos acerca de mi propia vida. No está siempre claro cuáles son estos deseos. Pero esto no es objeción contra la Teoría del Éxito. ¿Por qué debería estar siempre claro lo que vaya a ser mejor para mí?

Podemos recordar a continuación cómo difiere la Teoría del Éxito de la versión más amplia que hay de la Teoría Hedonista. Ambas teorías apelan a los deseos de una persona acerca de su propia vida. Pero los hedonistas apelan sólo a deseos que tratan de aquellos rasgos de nuestras vidas que son discernibles introspectivamente. Supongamos que mi deseo más intenso es resolver un problema científico. Los hedonistas afirman que será mejor para mí, para el resto de mi vida, creer que estoy resolviendo este problema. Según su modo de ver las cosas, no importará que mi creencia sea falsa, puesto que esto no supondrá ninguna diferencia en la calidad de mi experiencia vital. Conocer y creer falsamente no son diferentes experiencias. Según la Teoría del Éxito, será peor para mí que mi creencia sea falsa. Lo que quiero es resolver este problema. Será peor para mí que este deseo no se realice, aunque yo crea que se realiza.

Podemos recordar finalmente que hay dos versiones tanto de la Teoría Hedonista como de la del Éxito. Una versión apela a la suma total de la realización de los deseos *locales*, o particulares, de una persona. La otra versión apela sólo a los deseos *globales* de una persona: sus preferencias sobre partes de su vida, o bien sobre su vida entera. Yo podría preferir globalmente una de dos vidas posibles aunque ello implicase una suma total más pequeña de realización de deseo local. Una preferencia global como esta es la de las parejas que se arrojan a precipicios en la cumbre de su éxtasis.

Podríamos distinguir otras versiones de la Teoría de la Realización de los Deseos. Pero esto es innecesario aquí. Desafiaré a esos teóricos del Propio Interés que asumen alguna versión de la Teoría de la Realización de los Deseos. De nuevo usaré «deseo» para significar «deseo relevante», dado que diferentes versiones de esta teoría apelan a deseos diferentes. La mayor parte de mis observaciones se aplicarían a cualquiera de las versiones que se adopten.

Según la Teoría de la Realización de los Deseos, el axioma de la Benevolencia Racional de Sidgwick se convierte en

> (BR1) Lo que cada persona tiene más razón para hacer es lo que vaya a realizar de la mejor manera posible los deseos de todos,

y la teoría del Propio Interés se convierte en

> (PI11) Lo que cada persona tiene más razón para hacer es lo que vaya a realizar de la mejor manera posible, o le vaya a poner en disposición de realizar, la totalidad de sus propios deseos.

Un teórico del Propio Interés tiene que rechazar (BR1). Como señalé, puede afirmar

> (PI10) Puedo racionalmente ignorar deseos que no son míos.

Un teórico del fin Presente podría añadir

> (P5) Puedo racionalmente ignorar ahora deseos que no son míos ahora.

[Escribo «podría» porque (P5) podría ser rechazado a partir de la versión Crítica de P.] De acuerdo con la Apelación de la Relatividad Plena, si aceptamos (PI10) deberíamos aceptar también (P5). Como un teórico PI tiene que aceptar (PI10), pero no puede aceptar (P5), tiene que rechazar la Apelación a la Relatividad Plena. Esta apelación establece que las razones pueden ser relativas, no sólo a personas particulares, sino también a momentos temporales particulares. El teórico PI podría contestar que, mientras que tiene

un gran significado racional la cuestión de *quién* tiene determinado deseo, no lo tiene la cuestión de *cuándo* se tiene el deseo.

¿Es así? ¿Debería yo tratar de realizar mis deseos *pasados*? Una pregunta similar puede hacerse en relación con la Teoría de la Realización de los Deseos. ¿Es la realización de mis deseos pasados buena para mí, y su no realización mala para mí? A no ser los últimos deseos de los muertos, los deseos pasados raramente son discutidos por los teóricos de la realización de los deseos. Tal vez sea así porque la pregunta análoga no puede surgir en determinadas versiones de la Teoría Hedonista más antigua. Ahora no puedo mejorar la calidad de mis experiencias pasadas. Pero yo podría ser capaz de realizar mis deseos pasados aunque ya no quiera hacerlo. ¿Tengo una razón para hacerlo?

Algunos deseos están implícitamente condicionados a su propia persistencia. Si ahora quiero nadar cuando salga más tarde la luna, puedo querer hacerlo sólo si, cuando sale la luna, todavía quiero nadar. Si un deseo está condicionado a su propia persistencia, obviamente puede ser ignorado una vez que ha pasado.

Hay una clase de deseos muchos de los cuales están implícitamente condicionados de este modo. Estos son los deseos cuya realización pensamos que nos daría satisfacción, o cuya no realización pensamos que nos haría desgraciados. Nada de esto sucedería después de que hubiésemos dejado de tener estos deseos. Por eso es natural para muchos de estos deseos estar condicionados a su propia persistencia.

En el caso de otros deseos no hay tal razón general para que esto ocurra así. Supongamos que me encuentro con una desconocida en un tren. Ella me describe las ambiciones de su vida y las esperanzas y miedos con los que contempla sus posibilidades de éxito. Al final del viaje se ha despertado mi simpatía por ella, y deseo intensamente que esta desconocida tenga éxito. Tengo este intenso deseo aunque sé que nunca nos volveremos a ver. Este deseo no está implícitamente condicionado a su propia persistencia. Lo mismo ocurre con muchos otros deseos, de muchas clases.

Los ejemplos más claros son los deseos que algunos tienen en relación con lo que vaya a suceder después de que mueran. Supongamos que yo no creo que haya una vida de ultratumba. Como creo que mi muerte será mi extinción, mis deseos acerca de lo que ocurra después no pueden estar condicionados a su propia persistencia hasta el momento de la realización. Creo que esta condición no podría realizarse, y sin embargo sigo teniendo estos deseos. Y estos deseos serán, cuando yo esté muerto, deseos pasados que no estuvieron condicionados a su propia persistencia. Éstos —los deseos incondicionales de los muertos— no les causan problemas a algunos teóricos de la realización de los deseos. Ellos aceptan la tesis de que, al hacer que tales deseos no se realicen, obramos en contra de los intereses de los muertos.

Si asumen en cambio la Teoría del Éxito, no dirán esto sobre todos esos deseos. Uno de mis deseos más intensos es que Venecia nunca se destruya. Supongamos que cuando yo haya muerto una inundación destruye Venecia. Según la Teoría del Éxito, esto no iría en contra de mis intereses, ni implicaría que yo tuve una vida peor. Pero la Teoría del Éxito califica algunos sucesos como malos para alguien que está muerto. Supongamos que trabajo cincuenta años tratando de asegurar que Venecia se salvará. Según la Teoría del Éxito, será entonces peor para mí si, estando ya muerto, Venecia es destruida. Esto verificará que el trabajo de mi vida fue en vano. Cuando Venecia se destruya deberíamos afirmar que mi vida marchó peor de lo que antes habíamos pensado.

¿Deberíamos aceptar esta última afirmación? Parece justificable, pero también su negación. Si negamos esta tesis, parece que estamos apelando a la Teoría Hedonista. Parece que estamos asumiendo que no puede ser malo para mí que el trabajo de mi vida haya sido en vano, si yo nunca lo voy a saber. Es difícil ver por qué deberíamos creer esto, a no ser que asumamos que un suceso no puede ser malo para mí si no tiene como resultado una diferencia en la calidad de mi experiencia vital.

Aunque sean moralmente interesantes, los deseos de los muertos no son relevantes en esta discusión. Estoy considerando a los teóricos del Propio Interés que asumen alguna versión de la Teoría de la Realización de los Deseos. De acuerdo con ellos, lo que cada uno de nosotros tiene más razón para hacer es lo que vaya a reali-

zar mejor, o vaya a ponerle en condiciones de realizar, todos sus deseos a lo largo de toda su vida. No puedo preguntar si, cuando alguien está muerto, debería tratar de realizar sus deseos pasados. Tengo que hacer mi pregunta acerca de los deseos pasados de una persona viva. Tienen que ser deseos que no han persistido, pero que no estaban condicionados a su propia persistencia.

Podemos variar el mismo ejemplo. Supongamos que, durante cincuenta años, no sólo trabajo para tratar de salvar Venecia sino que también hago pagos regulares al Fondo para la Preservación de Venecia. A lo largo de estos cincuenta años mis dos deseos más intensos son que Venecia se salve, y que yo sea uno de sus salvadores. Estos deseos no están condicionados a su propia persistencia. Quiero que Venecia se salve, y que yo sea uno de sus salvadores, aunque más adelante deje de tener estos deseos.

Supongamos a continuación que dejo de tener estos deseos. Porque mis gustos en arquitectura cambian, dejo de preocuparme por el destino de la ciudad. ¿Todavía tengo una razón para contribuir al Fondo de Venecia? ¿Tengo tal razón según la versión de PI que apela a la Realización de los Deseos? Tengo una razón para dejar de contribuir, puesto que con el dinero que me ahorro podría realizar algunos de mis deseos presentes o futuros. ¿Tengo una razón contraria para seguir contribuyendo? Si sigo haciendo pagos, esto puede ayudar a realizar dos de mis deseos pasados. Aunque yo ya no tenga esos deseos, fueron mis deseos más intensos durante cincuenta años.

La pregunta de mayor alcance es esta. Según la Teoría de la Realización de los Deseos, ¿debería yo dar el mismo peso a *todos* mis deseos, pasados, presentes y futuros? ¿Debería dar la misma importancia a todos esos deseos cuando decida qué va a ser mejor para mí? (Quiero decir «la misma importancia, si no intervienen otros factores». Debería dar menos peso a uno de dos deseos si es más débil, o si me arrepiento de tenerlo, o si ciertas otras tesis son verdaderas. Pero ¿debería dar un peso menor a alguno de mis deseos porque no son mis deseos *presentes*?)

Supongamos que un teórico PI responde No. Y supongamos que también afirma que, al menos en el caso de deseos muy inten-

sos, su realización vale más si duran más tiempo. Si yo dejara de contribuir, esto me permitiría realizar algunos de mis deseos presentes y futuros. Pero sólo viviré unos pocos años más. De modo que puede suceder que, si tenemos en cuenta tanto la fuerza como la duración, mis deseos presentes y futuros, juntos, valdrían menos que los que fueron mis dos deseos más intensos durante cincuenta años. Si a todos mis deseos a lo largo de mi vida se les debiera dar igual peso, el teórico PI puede tener que concluir que sería peor para mí que yo dejara ahora de contribuir. Tiene que afirmar entonces que sería para mí irracional hacerlo. Sería irracional dejar de contribuir aunque yo ahora no tengo, y nunca tendré más adelante, *ningún* deseo de contribuir.

Esta conclusión puede violentar al teórico del Propio Interés. Quizás esté tentado a conceder que un agente racional puede ignorar sus deseos pasados. Pero, si la razón de esta tesis es que estos deseos pertenecen al pasado, puede que esta sea una concesión dañina. El teórico PI tiene entonces que abandonar la tesis de que no puede ser racionalmente significativo *cuándo* se tiene determinado deseo. Pero todavía tiene que afirmar que un agente racional debería dar el mismo peso a sus deseos presentes y futuros. Y si esta tesis no puede ser apoyada por una apelación a la neutralidad temporal, puede que sea más difícil de defender.

60. Deseos que dependen de juicios de valor o ideales

Si el teórico PI quiere apelar a la neutralidad temporal, tiene que dar alguna otra razón por la que podamos ignorar nuestros deseos pasados. Podría apelar a uno de los modos en que podemos perder un deseo. Podemos cambiar de opinión. En cierto sentido, cualquier cambio en nuestros deseos implica un cambio de opinión. A lo que aquí nos referimos es a un cambio en algún juicio de valor, o ideal: un cambio de opinión acerca de lo que vale la pena desear.

Distinguí entre lo que tengo más razón para hacer y lo que, dadas mis creencias, sería racional para mí hacer. Si me han envenenado el vino, beberlo no es lo que tengo más razón para hacer. Pero

si no tengo razón para pensar que lo han envenenado, no obraría irracionalmente si me lo bebiera. Mi pregunta principal es sobre lo que tenemos más razón para hacer. Pero lo que sigue es sobre lo que, dadas las creencias de alguien, sería racional que hiciese. Hay quienes llaman a ésta la cuestión de lo que es *subjetivamente* racional.

El teórico PI puede decir: «Cualquiera puede racionalmente ignorar los deseos que perdió porque cambió de opinión. Y tienes que haber cambiado de opinión cuando dejaste de querer que Venecia se salvase. En contraste, cuando muere alguien, sus deseos se vuelven pasados sin un cambio de opinión. Por eso, cuando consideramos los intereses de una persona no deberíamos ignorar los deseos que tenía a la hora de morir».

Esto no puede ser una respuesta completa, puesto que incluye sólo a esos deseos que dependen de juicios de valor o ideales. Tenemos muchos deseos más simples, cuya pérdida no implica un cambio de opinión.

Cuando se aplica a casos en los que cambiamos de opinión, ¿es esta una buena respuesta? En un pequeño número de casos, discutidos después, consideramos nuestro cambio de opinión como una perversión. En todos los demás casos es plausible afirmar que, cuando alguien pierde un deseo a consecuencia de haber cambiado de opinión, puede racionalmente ignorar este deseo. Esto ocurre porque este deseo dependía de un juicio de valor o ideal que ahora rechaza.

Aunque esta afirmación resulta plausible, *no* es una buena respuesta en defensa de PI. Una afirmación similar se aplica a esos deseos que *después* serán producidos por un cambio de opinión. Supongamos que predigo que después tendré deseos que dependerán de juicios de valor o ideales que ahora rechazo. ¿Debería yo dar a esos deseos futuros el mismo peso que doy a mis deseos presentes? El teórico PI tiene que contestar Sí.

Nagel describe un ejemplo relevante:

> «Supongamos [que alguien] cree ahora que dentro de veinte años valorará la seguridad, el estatus, la riqueza y la tranquilidad, mientras que ahora valora el sexo, la espontaneidad, los riesgos frecuentes y las emociones fuertes. Una respuesta decidida a esta situación podría asumir una de dos formas. El individuo puede estar lo suficientemente convencido de la falta de valor de sus valores futuros inevitables, simplemente para rehusarles toda pretensión sobre su interés presente... Por otro lado, puede tratar tanto sus valores presentes como los futuros como preferencias, considerando a cada una de ellas como una fuente de razones bajo un principio más alto: "Vive con el estilo de vida de tu elección". Esto exigiría de él una cierta prudencia en lo que respecta a mantener abiertas las sendas a la respetabilidad final» [18].

Como escribe Nagel, PI requiere que este joven trate sus valores e ideales como si fueran meras preferencias. Sólo así podría ser racional para él dar igual peso a sus previsibles valores futuros. PI afirma que no tiene que actuar de modos que, previsiblemente, vaya a lamentar en el futuro. Así que no tiene que dar su apoyo a movimientos políticos, ni firmar peticiones, si esto pudiera dificultar o restringir seriamente las oportunidades de su más conservador Yo de madurez. Más exactamente, él puede obrar de esos modos, si un cálculo imparcial sale bien. En este cálculo puede descontar la menor probabilidad de que más tarde vaya a tener valores o ideales diferentes. Pero no tiene que descontar tales valores o ideales previsibles simplemente porque ahora piense que carecen de valor o son despreciables.

Si creemos que nuestros valores o ideales *no* son meras preferencias, y pueden justificarse más o menos, no podemos aceptar esta última tesis. No podemos aceptar completamente PI. Tenemos que tratar nuestros valores o ideales como P dice que debiéramos. Tenemos que darle un estatus especial a lo que *ahora* pensamos que está mejor justificado.

Este punto resulta aún más claro si consideramos deseos que, a diferencia de los que figuraban en el ejemplo de Nagel, descansan en creencias morales. Resulta más claro todavía si asumimos que estas creencias pueden ser, no sólo más o menos defendibles, sino lisa y llanamente verdaderas. El tema es entonces uno muy familiar

[18] Nagel (1), p. 74.

para todas las creencias. No podemos decir honestamente, «Q es verdadero, pero yo ahora no creo en Q». Ahora tenemos que creer que es verdadero lo que ahora creemos que es verdadero. Pero esta afirmación sobre nuestras creencias no incluye a nuestras creencias pasadas o futuras. Podemos decir honestamente, «Q es verdadero, pero yo no tenía la costumbre de creer en Q, y puedo dejar de creer en Q en el futuro».

El punto correspondiente acerca de los deseos evaluativos necesita ser formulado cuidadosamente. Nagel escribe: «El individuo puede estar lo suficientemente convencido de la falta de valor de sus valores futuros inevitables para rehusarles toda pretensión en lo que hace a su interés presente. Él entonces consideraría sus valores presentes válidos también para el futuro, y no se derivarían razones prudenciales de sus concepciones futuras esperadas». Nagel añade que, si es así como el individuo responde, «su posición se podría formular en términos de razones intemporales». Sus razones se formularían de una forma *intemporal* porque apelarían a valores que él considera válidos intemporalmente. Pero con todo sería cierto que, en otro sentido, sus razones serían tanto *relativas al momento temporal* como *relativas al agente*. La persona en cuestión trata de obrar sobre la base de los valores que *ahora considera* válidos. Esto es lo que le dice que haga la teoría del fin Presente.

Según la teoría del Propio Interés, este joven tiene que dar el mismo peso a sus valores e ideales presentes y a sus valores e ideales futuros previstos. Esto significaría dar el mismo peso a lo que él ahora considera justificado y a lo que él ahora considera sin valor o despreciable. Esto es claramente irracional. Puede incluso ser lógicamente imposible.

Hemos alcanzado una conclusión general. De acuerdo con PI, yo debería dar el mismo peso a todos mis deseos, presentes y futuros. Esta tesis se aplica incluso a los deseos futuros que dependerán de un cambio en mis juicios de valor o ideales. Cuando se aplica a estos deseos, esta tesis es indefendible. En el caso de razones para actuar que se basan en juicios de valor, o ideales, un agente racional tiene que dar prioridad a los valores o ideales que acepta ahora. *En el caso de estas razones, la teoría correcta no es PI sino P.*

Hay razones ulteriores para esta conclusión. Supongamos que creo que, con un conocimiento y una experiencia en aumento, me iré haciendo más sabio. Según esta asunción, debería dar a mis deseos evaluativos *futuros más* peso que el que doy a mis deseos evaluativos presentes, puesto que mis deseos futuros estarán mejor justificados. Esta tesis puede parecer que entra en conflicto con P. Pero no es así. Si yo no sólo asumo que siempre me estoy haciendo más sabio sino que además puedo predecir ahora algún futuro cambio de opinión particular, ya he, de hecho, cambiado de opinión. Si ahora creo que determinada creencia posterior estará mejor justificada, debería tener esta creencia ahora. De modo que la asunción de que me estoy haciendo más sabio no supone ninguna objeción contra P. Incluso según esta asunción, puedo dar todavía un estatus especial a lo que ahora considero justificado.

La asunción contraria es que, a medida que transcurre el tiempo, me haré menos sabio, y que el cambio de mis ideales será una perversión. La pérdida de ideales es un lugar común; y el juicio muchas veces se pierde. (En ediciones sucesivas de sus *Poemas Selectos* muchos poetas hacen selecciones cada vez peores.) Según esta asunción, yo debería dar más peso a lo que acostumbraba a valorar o a creer. Pero de la misma manera, si acepto esta asunción, todavía creería lo que acostumbraba a creer, aunque me importara menos.

De estas asunciones contrarias, ninguna parece en resumidas cuentas con más probabilidades de ser justificada. Ya que están en conflicto, podemos sugerir, como compromiso, la neutralidad temporal que la teoría del Propio Interés requiere. Y hay un argumento escéptico que puede que parezca favorecer tal neutralidad. Puede que yo esté impresionado por la presunción de mi certeza presente. ¿Por qué debería yo asumir que es más probable que yo tenga razón *ahora*?

Aunque dé apoyo a un cambio hacia la neutralidad temporal, este argumento no puede ayudar al teórico del Propio Interés. Como otros argumentos, nos lleva más allá de su teoría. ¿Por qué debería yo asumir que es más probable que *yo* tenga razón? El argumento escéptico cuestiona mi confianza no sólo en actuar según los valores o ideales que acepto *ahora*, sino además en actuar según los

valores o ideales que *yo* acepto. El argumento apoya el dar igual peso a todos los valores o ideales rivales de todas las personas de las que considero probable que tengan tanta razón como yo.

Si asumimos que nuestros valores o ideales pueden estar peor o mejor justificados, es una cuestión desconcertante cómo deberíamos reaccionar a este argumento escéptico. ¿No es arrogante y además irracional asumir que los valores o los ideales mejor justificados son *los míos*? Aquí puedo ser empujado en ambas direcciones. Esta asunción puede en verdad parecer arrogante, o irracional. Pero puede que sea absurdo afirmar que no tengo que valorar más lo que valoro más. Esto es como la afirmación de que no tengo que creer que es verdadero lo que creo que es verdadero. Y esta afirmación puede que socave mi creencia en cualquiera de estos valores [19].

Hay argumentos en las dos direcciones. Pero ningún argumento apoya a PI. Una vez más, PI ocupa una posición intermedia injustificable. Si el argumento escéptico tiene éxito, la Neutralidad gana. Como «el liberal» de Hare, yo debería dar igual peso a los valores e ideales de toda persona bien informada y racional [20]. Supongamos que este argumento fracasa. Si yo debiese dar más peso a *mis* valores e ideales, también debería dar más peso *ahora* a lo que valoro o creo *ahora*. El argumento para la primera tesis, cuando se lleva hasta el final, justifica la última.

El conflicto entre estos argumentos es un ejemplo de lo que Nagel llama el conflicto entre lo *subjetivo* y lo *objetivo* [21]. Las tesis de la objetividad son, aquí, las tesis de la *intersubjetividad*. Me llevan

[19] Cf. Williams (3), p. 215:

«... cosas tales como el profundo cariño hacia otras personas... no pueden incorporar al mismo tiempo la concepción imparcial, y... también corren el riesgo de infringirla. Corren ese riesgo si en absoluto existen. Pero a no ser que estas cosas existan, no habrá suficiente sustancia ni convicción en la vida de un hombre para imponerle lealtad a la vida misma. La vida tiene que tener sustancia si algo va a tener sentido, incluyendo la adhesión al sistema imparcial. Pero si tiene sustancia, entonces no le puede garantizar al sistema imparcial la importancia suprema, y el ajuste en ella de ese sistema será, en el límite, inseguro».

[20] Véase Hare (4), pp. 159-67, y 175-84.

[21] Nagel (4), capt. 14.

más allá de los límites de mi propia vida, y me dicen que les dé peso a los valores o ideales de otros. Y las tesis de la subjetividad son las tesis de lo que yo *ahora* creo. Son las tesis de mi punto de vista presente. Cuando estamos interesados en valores o ideales, no podemos justificadamente afirmar que la unidad significativa es la vida entera de uno. No podemos afirmar que yo debería dar ahora el mismo peso a todos y sólo los valores o ideales que, en cualquier momento temporal, tuve o tengo o tendré. Una vez más, no podemos requerir neutralidad temporal y a la vez rechazar el requisito de neutralidad interpersonal. *Los dos* argumentos que acabamos de detallar cuentan contra la teoría del Propio Interés.

61. Simples deseos pasados

Estas afirmaciones se aplican únicamente a deseos cuya pérdida implica un cambio de opinión. Hay incontables deseos con los que esto no pasa. Esto es así aunque, como algunos afirman, todos los deseos impliquen una evaluación. Hay una vasta gama de posibles objetos de interés valiosos. Dados los límites de nuestras mentes y de nuestras vidas, cada uno de nosotros puede interesarse intensamente tan sólo en unos pocos de estos objetos. Y nuestro interés puede ir de uno a otro de ellos sin tener que creer que aquello por lo que estamos interesados ahora es más digno de interés. Esto es claro en el caso en que, al final de mi viaje en tren, tengo el intenso deseo de que la desconocida tenga éxito. Más tarde perderé este deseo, pero esto *no* será porque decida que importa menos el que esta desconocida tenga éxito.

Aquí tenemos otro caso. Cuando yo era joven lo que más deseaba era ser poeta. Este deseo no estaba condicionado a su propia persistencia. Yo no quería ser poeta sólo si esto siguiera siendo más adelante lo que yo quería. Ahora que soy más mayor ya no tengo ese deseo. He cambiado de opinión en el sentido más restringido de que he cambiado de intenciones. Pero no he decidido que la poesía sea de ningún modo menos importante o menos admirable. ¿Mi deseo pasado me da una razón para tratar de escribir poemas ahora, aunque ahora no tenga el deseo de hacerlo?

Como mi pérdida de este deseo no llevó consigo cambio alguno en mis juicios de valor, el teórico del Propio Interés sólo tiene las siguientes alternativas. Podría mantener la tesis de que un agente racional debería ser temporalmente neutral: de que él debería tratar de hacer lo que fuera a realizar mejor todos sus deseos a lo largo de su vida entera. Entonces tendría que aceptar que, según su teoría, un agente racional debe dar peso a aquellos de sus deseos pasados cuya pérdida no llevara consigo un cambio en sus juicios de valor. Las únicas excepciones son esos deseos pasados que estaban condicionados a su propia persistencia. Tendría por tanto que afirmar que yo tengo una poderosa razón para tratar de escribir poemas ahora, porque esto fue lo que más quise durante muchos años. Tengo una poderosa razón para tratar de escribir poemas, aunque ya no tenga el más mínimo deseo de hacerlo. La mayoría de nosotros encontraría esta afirmación difícil de creer.

Si el teórico del Propio Interés está de acuerdo, tiene que rechazar el requisito de neutralidad temporal. Tiene que afirmar que no es irracional no dar peso a los propios deseos pasados. Pero una vez que ha abandonado la neutralidad temporal, o la tesis de que no puede ser racionalmente significativo *cuándo* tenemos determinado deseo, el resto de su teoría es más difícil de defender. Todavía tiene que insistir en la neutralidad temporal que se establecería entre nuestros deseos presentes y la totalidad de nuestros deseos futuros. Tiene que insistir en que yo debería dar igual peso a todo lo que *vaya a* querer previsiblemente en el futuro, aunque ahora no lo quiera. Esto es más difícil de justificar si a los deseos que yo *tuve* no se les puede dar ningún peso. Si puede tener significación racional que los deseos sean pasados, ¿por qué no puede tener significación racional el que no sean presentes? ¿Por qué debería ser tratado el presente como si fuera sólo parte del futuro? Si yo puedo *no* dar peso a lo que yo deseaba, *porque* no lo deseo ahora, ¿por qué tengo que dar *igual* peso a lo que desearé, cuando no lo deseo ahora?

En las tres últimas secciones he tratado a esos teóricos PI que hacen suya alguna versión de la Teoría de la Realización de Deseos acerca del propio interés. He defendido dos conclusiones. Algunos

de nuestros deseos se basan en juicios de valor, o ideales, o creencias morales. En el caso de estos deseos, tenemos que aceptar P antes que PI. En el caso de nuestros otros deseos, un teórico PI tiene dos alternativas. Podría insistir en que deberíamos tratar de realizar nuestros deseos pasados incondicionales: los deseos incondicionales que una vez tuvimos pero que ya no tenemos. Esto es difícil de creer. La otra alternativa del teórico PI es abandonar la afirmación de que no puede suponer diferencia alguna *cuándo* se tiene un deseo. Si abandona esta tesis, el teórico PI necesitará una nueva explicación de por qué deberíamos dar el mismo peso a nuestros deseos presentes y a nuestros deseos futuros. Y puede ser difícil encontrar esta nueva explicación.

Si el teórico PI abandona la apelación a la neutralidad temporal, entonces tendrá que abandonar lo que llamé su Segunda Respuesta. Según esta nueva concepción, yo *tuve* razones para tratar de realizar mis deseos pasados; pero, como ya no los tengo, tampoco tengo razones para tratar de realizarlos ahora. Si esto es así, tendrá que renunciar a su tesis de que la fuerza de cualquier razón se extiende a través del tiempo.

Mi otra conclusión también socava esta tesis. Tengo razones para actuar que me las proporcionan mis valores, o ideales, o creencias morales, presentes. Si más adelante cambio de opinión, tendré razones contrarias para actuar. En el caso de estas razones, tenemos que aceptar P antes que PI. La fuerza de *estas* razones no se extiende a lo largo del tiempo.

62. ¿ES IRRACIONAL PREOCUPARSE MENOS POR NUESTRO FUTURO MÁS LEJANO?

El teórico PI no puede afirmar que la fuerza de cualquier razón se extienda a través del tiempo. Por eso tiene que apelar a su otro argumento. Tiene que afirmar que la inclinación a favor de uno mismo es supremamente racional. Compararé ahora esta inclinación con otro patrón de intereses corriente: preocuparse menos por nuestro futuro más lejano. Como este es el blanco favorito de los

teóricos PI, voy a cuestionar su teoría allí donde creen que es más plausible.

Volveré también de lo que es distintivo de la Teoría de la Realización de Deseos a lo que es común a todas las teorías plausibles del propio interés. De acuerdo con todas estas teorías, la Teoría Hedonista es, como mínimo, parte de la verdad. Parte de lo que hace que nuestras vidas marchen mejor es el goce, la felicidad y la evitación del dolor y el sufrimiento. Estas cosas serán lo que importe en los casos que voy a discutir. Elijo estos casos en parte porque son simples, y en parte porque son los casos en que la teoría del Propio Interés parece a muchos máximamente convincente.

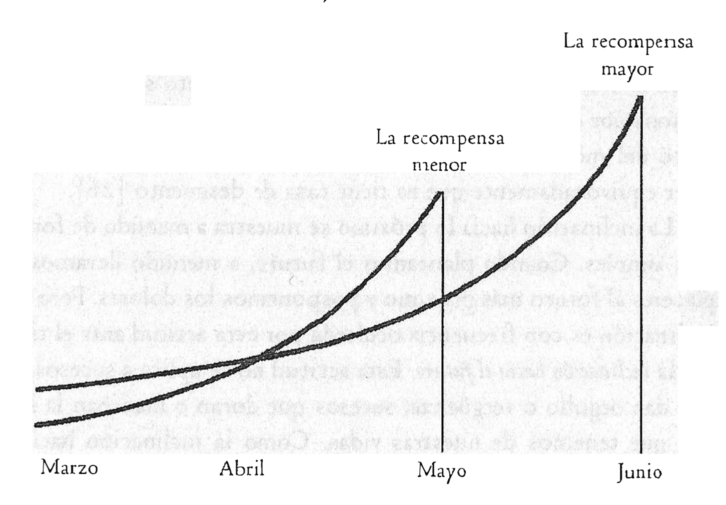
Bentham afirmó que, al decidir el valor de cualquier placer futuro, deberíamos considerar cuán *pronto* lo disfrutaremos [22]. C. I. Lewis sugiere que esto puede haber sido un vago recordatorio de que «los placeres más cercanos son en general los más seguros» [23]. Pero la afirmación sería entonces redundante, porque Bentham nos aconseja directamente que consideremos la probabilidad de los placeres futuros. Si tomamos su afirmación en sentido estricto, nos dice que prefiramos los placeres más cercanos justo porque son más cercanos. Lo cual compromete a Bentham con la idea de que, «aunque deberíamos estar racionalmente interesados en el futuro, debiéramos estar menos interesados en él en la medida en que sea más remoto —y esto con total independencia de cualquier duda que se adjunte a lo más remoto»—. Lewis denomina a esto «el principio de la prudencia fraccionaria». Como él admite, «expresa una actitud que los humanos tienden a adoptar». Pero lo considera tan claramente irracional que según él no vale la pena discutirlo.

Llamo a esta actitud la inclinación hacia lo próximo. Hume describe uno de los modos en que se revela esta inclinación: «Al reflexionar sobre una acción que realizaré dentro de doce meses prefiero siempre el bien mayor, sin importarme si en ese momento estará más o menos próximo... Pero a medida que me voy aproximando... surge una nueva inclinación hacia el bien presente, y me resulta difícil adherirme inflexiblemente a mi propósito y resolución primeros» [24].

[22] Bentham, capt. IV, segundo párrafo.
[23] Lewis (1), p. 483.

Las palabras de Hume sugieren que esta inclinación se aplica sólo al futuro inmediato. Pero una descripción más ajustada es esta. Tenemos una *tasa de descuento* con respecto al tiempo, y descontamos el futuro más próximo a una razón mayor. Por eso no nos «adherimos» a nuestras «resoluciones». Aquí van dos ejemplos. Tomo la decisión de que cuando, dentro de cinco minutos, me quite el esparadrapo de la pierna, lo arrancaré de un golpe, prefiriendo ahora la perspectiva de un dolor agudo que dura un momento a la larga molestia de quitarme el esparadrapo pelo a pelo. Pero cuando llega el momento invierto mi decisión. De forma parecida, tomo la decisión de que cuando dentro de cinco años comience mi carrera profesional, pasaré la primera parte de la misma en un puesto tedioso pero con la probabilidad, en la segunda parte, de llevarme a la cima. Pero cuando llega el momento invierto de nuevo mi decisión. En ambos casos, contemplada la situación desde la distancia, parece que vale la pena sufrir algo malo para conseguir el bien que viene a continuación. Pero, cuando ambos están más cerca, la escala se inclina en el otro sentido.

Otro caso de muestra abajo.



[24] Hume (1), Libro III, Parte II, Sección VII.

La altura de cada curva muestra cuánto me interesa en cada momento una de las dos posibles recompensas futuras. Me preocupa menos el futuro más lejano; y me preocupa tanto menos en el futuro más próximo. Esto lo pone de manifiesto el hecho de que estas curvas sean más empinadas justo antes de que consiga estas recompensas. Puede ser útil reformular estas tesis. Si uno de dos sucesos similares va a ocurrir un mes más tarde, ahora me preocupa menos. Mi interés será alguna porción de mi interés por el otro suceso de un mes antes. Pero en caso de que estos dos sucesos se hallen *más lejos* en el futuro, habrá *menos* diferencia proporcionada en el interés que en el presente tengo por ellos. Y la diferencia proporcionada será *la más grande* cuando el suceso anterior esté en el futuro inmediato.

Estas afirmaciones explican por qué, en mi diagrama, las dos curvas se cruzan. Cuando se cruzan, mi preferencia cambia. Juzgando desde marzo, prefiero la recompensa mayor en junio a la recompensa menor en mayo. Juzgando desde el final de abril, prefiero la recompensa menor en mayo [25].

Consideremos a continuación a alguien con un diferente tipo de tasa de descuento: una que es *exponencial*. Tal persona descuenta el futuro a una razón constante de *n* por ciento por mes. Siempre habrá la *misma* diferencia proporcionada en cuánto se interesa esta persona por dos sucesos futuros. Sus preferencias no cambiarán por tanto del modo descrito arriba. Como esto es así, puede llegar a creer equivocadamente que *no* tiene tasa de descuento [26].

La inclinación hacia lo próximo se muestra a menudo de formas más simples. Cuando planeamos el futuro, a menudo llevamos los placeres al futuro más próximo y posponemos los dolores. Pero esta inclinación es con frecuencia ocultada por otra actitud ante el tiempo, la *inclinación hacia el futuro*. Esta actitud no se aplica a sucesos que nos dan orgullo o vergüenza: sucesos que doran o manchan la imagen que tenemos de nuestras vidas. Como la inclinación hacia lo próximo, la inclinación hacia el futuro se aplica más claramente a

[25] Tomo este diagrama, a través de R. Nozick, de G. Ainslie.
[26] Véase Strotz.

sucesos que son en sí mismos agradables o penosos. Pensar en tales sucesos nos afecta más cuando se hallan en el futuro que cuando se hallan en el pasado. Ver en el futuro un placer es, en general, más agradable que verlo en el pasado. Y en el caso de los dolores la diferencia es aún mayor. Comparemos los estados mentales de un colegial inglés antes y después de una paliza.

Con frecuencia obramos de formas que pueden dar la impresión de que *no* tenemos ninguna inclinación hacia lo próximo: llevamos los *dolores* al futuro más próximo y posponemos los placeres. La inclinación hacia el futuro nos proporciona la explicación. Queremos dejar los dolores detrás de nosotros y mantener los placeres ante nosotros. Como la segunda inclinación contrarresta la primera, nuestra tendencia a actuar en estas dos formas no puede mostrar que no tengamos inclinación alguna hacia lo próximo. Nuestra inclinación hacia lo próximo puede estar siempre compensada por nuestra inclinación hacia el futuro. Recuerdo haber decidido, tras soplar las diez velas de mi pastel de cumpleaños, que en el futuro siempre me comería el mazapán al final y no al principio.

Aquí tenemos otro ejemplo. Supongamos que tengo que elegir cuándo voy a sufrir un tratamiento doloroso. Si espero un año, hasta que el hospital tenga un equipo nuevo, el tratamiento sólo será la mitad de doloroso. Y supongamos que mi tasa de descuento se reduce a la mitad en el espacio de un año. Si pospongo el tratamiento un año, ahora me preocupará sólo una cuarta parte. Será en sí mismo la mitad de doloroso, y ahora lo descuento por la mitad. Pero si pospongo el tratamiento tendré una anticipación penosa que durará un año entero. Esta perspectiva, aunque descontada, puede parecerme peor ahora que la perspectiva del tratamiento inmediato. Así las cosas, a pesar de mi inclinación hacia lo próximo, elegiré pasar por el tratamiento ahora, cuando de hecho va a ser dos veces más doloroso.

Hay algunas personas a las que no les importa más lo que está cerca. Algunas incluso se preocupan más por lo remoto. La propensión a ahorrar, o a posponer la gratificación, puede ser compulsiva. Pero no tenemos necesidad de decidir aquí cuántas personas tienen la inclinación hacia lo próximo.

Hay otra cuestión más importante. A menudo se afirma que esta inclinación está siempre causada por algún fallo de la imaginación, o alguna creencia falsa. Se dice, por ejemplo, que cuando imaginamos dolores en el futuro más lejano, los imaginamos con menos viveza, o creemos confusamente que de algún modo serán menos reales, o menos dolorosos. Desde que Platón hiciera esta afirmación [27], se ha repetido muy a menudo. Por eso Pigou declaró que tenemos «un defecto en la facultad telescópica» [28]. Y esta afirmación está incrustada en nuestro lenguaje. Alguien a quien le importa menos su futuro lejano es *imprudente*, o *imprevisor* —palabras que significan en latín que no ve el futuro—. Y algunos economistas llaman a esta actitud *miopía*.

La afirmación de Platón es a menudo verdadera. En el caso de muchas personas, nos proporciona una explicación parcial de su inclinación hacia lo próximo. Sería importante el que esta tesis fuera verdadera siempre, y nos aportara toda la explicación. Si así fuese, esta inclinación nunca sobreviviría al proceso de *deliberación ideal*. Lo cual reduciría el monto de conflicto entre PI y algunas versiones de P. Pero, como sostengo en la Sección 73 y en el Apéndice C, PI y P, aun así, no coincidirían.

Aunque se trate de una cuestión fáctica, estoy totalmente seguro de que la tesis de Platón es con frecuencia falsa. Una prueba sería esta. En cierto experimento alguien tiene que decidir si acepta sufrir un dolor ante la perspectiva de un placer. Esta persona sabe que, cuando haya tomado la decisión, tomará una píldora que hará que olvide la decisión, lo que hará irrelevante los placeres o los sufrimientos de la anticipación. La persona sabe también que no le diremos nada acerca del tiempo fijado para el dolor y para el placer hasta el momento antes de que tome la decisión. Describimos minuciosamente lo que implicarían las dos experiencias. Para que pueda tomar una decisión plenamente informada, la persona se pone a imaginar tan vívidamente como le es posible cómo sería sufrir el dolor y disfrutar el placer.

[27] Platón (1)
[28] A. C. Pigou, *The Economics of Welfare* (*La economía del bienestar*), Londres, Macmillan, 1932, p. 23.

Entonces le decimos que el dolor sería inmediato y el placer se pospondría un año. ¿Acaso el placer le parecería ahora menos vívido? Al menos en mi propio caso, tengo la seguridad de que no. Supongamos que, si el dolor fuese inmediato y el placer se pospusiera un año, la persona tendría una ligera preferencia por no tener ninguno de los dos. Toma la decisión de que este placer no es lo bastante grande como para hacer que valga la pena sufrir el dolor. Entonces le decimos que estábamos mal informados: el placer sería inmediato, y el dolor se pospondría un año. Creo probable que la preferencia de esta persona ahora cambiaría. Ahora podría decidir que vale la pena soportar el dolor con vistas a conseguir el placer. Puesto que la persona imaginó las dos experiencias cuando no sabía nada del momento temporal que se les había fijado, este cambio de sus preferencias no sería producido por el supuesto hecho de que las experiencias posteriores siempre parecen menos vívidas en la imaginación. Tendríamos una buena razón para pensar que la persona en cuestión tiene una inclinación hacia lo próximo, y la tiene de un modo que sobrevive a la deliberación ideal.

Creo que la inclinación hacia lo próximo es común. Pero, para evitar la polémica, podemos hablar de una persona imaginaria. Esta persona se interesa más por su futuro más próximo, simplemente porque es más próximo; y lo hace así aunque conozca los hechos y piense con claridad. Llamaré a esta persona *Próximus*. No afectará para nada a la argumentación el que, como pienso, haya muchas personas reales que son así.

Muchas veces no está claro lo que sería mejor para alguien, o lo que más favorecería sus intereses, tanto porque los hechos son dudosos como a causa del desacuerdo entre las teorías rivales del propio interés. Pero según todas las teorías plausibles hay una cuestión sobre la que reina el acuerdo. Cuando estamos decidiendo lo que favorece los intereses de alguien, deberíamos descontar la incertidumbre, pero no la mera lejanía temporal. No deberíamos dar un peso menor al futuro lejano de la persona, o dar mayor peso a sus deseos presentes.

Puede que creamos que a veces debiéramos dar un mayor peso a los deseos presentes de alguien. Puede que pensemos que es erróneo

ignorar los deseos presentes de la persona, forzándola a hacer lo que será mejor para ella en el futuro. Pero pensaríamos que esto es erróneo porque infringe la autonomía personal. Cosa que es consistente con mi tesis de que, según todas las teorías plausibles del propio interés, al decidir lo que sería lo mejor para alguien, *no* deberíamos dar mayor peso a los deseos presentes de la persona. Deberíamos dar igual peso a todas las partes de la vida de la persona.

Según la teoría del Propio Interés, alguien obra irracionalmente cuando hace lo que sabe que será peor para él. Mi hombre imaginario con frecuencia obra así. Como está inclinado hacia lo próximo, Próximus a menudo pospone los dolores deliberadamente, con el coste previsible de hacerlos peores. En estos casos, hace lo que sabe será peor para él.

Comparemos ahora su actitud con la de un hombre guiado por su propio interés. ¿Es la inclinación hacia lo próximo menos racional que la inclinación en favor de uno mismo? Es esencial para la defensa de PI que respondamos Sí.

Próximus conoce los hechos y piensa con claridad. Deberíamos añadir una asunción más. Los que tienen alguna inclinación puede que deseen estar libres de ella. Esto es muy común en el caso de la inclinación hacia lo próximo. Tras describir cómo esta inclinación le hace obrar en contra de sus intereses, Hume escribió, «yo lamento muchísimo esta debilidad natural» [30]. Deberíamos asumir que Próximus no se arrepiente en absoluto de ella. Sólo esta asunción hará justa nuestra comparación, porque los que se guían por el propio interés se supone comúnmente que no lamentan su inclinación en favor de sí mismos. Además, al rechazar PI, Próximus apelará a P; y la versión Crítica de P puede descontar deseos que el agente desea no haber tenido. Puede afirmarse que debemos dar a esos deseos menos peso, o incluso ninguno. Por ello deberíamos asumir que Próximus no lamenta su inclinación hacia lo próximo.

Puede objetarse que, cuando sufre los efectos de esta inclinación, tiene que arrepentirse de tenerla. Pero, como explico en la Sección 71, Próximus jamás lamenta ni tener esta inclinación ahora,

[30] Hume (1), Libro III, Parte II, Sección VII.

ni haberla tenido en el pasado más próximo. Como mucho se arrepiente de haberla tenido en el pasado lejano. Actúa como actúa a causa de su inclinación presente; y nunca lamenta *esta* inclinación.

63. Un argumento suicida

¿Cómo debería criticar a Próximus un teórico del Propio Interés? Enfrentado a la elección de un dolor moderado ahora, o un dolor mucho peor más tarde, Próximus con frecuencia elige deliberadamente el dolor peor. Y con frecuencia prefiere un placer pequeño pronto a un placer muyo mayor más tarde. Por eso tiene que afirmar, con Hume, que «tampoco es contrario a la razón preferir incluso un bien menor, aunque lo reconozca como tal, a otro mayor» [31]. Esto —la elección deliberada de lo que una persona admite que será peor para sí misma— puede parecer el caso de irracionalidad más claro posible. Un teórico PI podría decir, «La primera regla de la racionalidad es rechazar lo que sabes que es peor».

Próximus podría contestar: «Si la única diferencia entre dos dolores es que uno sería peor, acepto tu regla. Pero, en los casos que estamos discutiendo, hay otra diferencia. Cuando elijo entre dos dolores, considero no sólo lo dolorosos que serían sino además lo pronto que tendría que sufrirlos. No estoy simplemente eligiendo lo que sé que es peor. Elijo el peor de dos dolores únicamente cuando la proporción en la que es peor es, para mí ahora, compensada por la proporción en la cual está más lejos en el futuro».

El teórico PI tiene que contestar que es irracional tomar en cuenta la proximidad. De hecho podría afirmar, citando a Rawls, que «la mera posición temporal, o la distancia desde el presente, no es una razón para favorecer a un momento temporal sobre otro» [32]. Más en general, el teórico PI podría recuperar el requisito de la neutralidad temporal.

Defendí antes que, si el teórico PI asume la Teoría de la Realización de los Deseos sobre el propio interés, debería abando-

[31] Hume (1), Libro II, Parte III, Sección III.
[32] Rawls, p. 420.

nar el requisito de la neutralidad temporal. Según esta teoría del propio interés, este requisito implica, de manera no plausible, que deberíamos tratar de realizar ahora algunos de los deseos que una vez tuvimos, aunque ni ahora tengamos ni nunca más tarde vayamos a tener estos deseos. En lo que sigue, asumiré que el teórico PI rechaza la Teoría de la Realización de los Deseos, aceptando o bien la Teoría Hedonista, o bien alguna versión de la Teoría de la Lista Objetiva. Según esta asunción, si el teórico PI exige neutralidad temporal, no necesita afirmar que deberíamos tratar de realizar tales deseos pasados. Su tesis puede ser que, cuando estamos considerando placeres y dolores, o felicidad y sufrimiento, meras diferencias en disposición temporal no pueden tener significación racional.

El teórico PI ¿cómo debería dar apoyo a esta tesis? ¿Por qué no debería tomarse el tiempo en consideración? Podría decir:

> Una mera diferencia respecto a *cuando* ocurre algo no es una diferencia en su cualidad. El hecho de que un dolor está más lejos en el futuro no hará que, cuando llegue, sea menos doloroso.

 Este es un argumento *excelente*. Es con mucho la mejor objeción a la inclinación hacia lo próximo. Pero el teórico PI no puede utilizar este argumento. Es una navaja de dos filos. El mismo argumento puede ser utilizado contra la Teoría del Propio Interés. Igual que Próximus toma en cuenta *cuándo* se siente un dolor, el teórico PI toma en cuenta *quién* lo sentirá. Y una mera diferencia en quién siente un dolor no es una diferencia en su cualidad. El hecho de que un dolor sea de otro no lo hace en absoluto menos doloroso.

El teórico PI toma en cuenta (1) lo malos que serían los dolores, y (2) quién los sentiría. Por eso a veces elige el peor de dos dolores. A veces elige un dolor peor para otro antes que un dolor más pequeño para sí mismo. (Puede parecer que siempre haría esta elección. Pero esto da por sentado que el teórico PI tiene que ser puramente egoísta. Como dije, esto es un error. Alguien que acepta PI puede querer a otras personas. Puede, por tanto, ser peor para él escapar a algún dolor menor al coste de imponer un dolor peor a alguien que quiere.)

Próximus toma en cuenta (1) lo malos que serían los dolores (2) quién los sentirá, y (3) cuándo se sentirán. Le puede decir al teórico PI, «Si tomas en cuenta quién sentirá un dolor, ¿por qué no puedo tomar en cuenta cuándo se siente un dolor?». Pueden darse respuestas a esta pregunta. Puede haber argumentos que muestren que las diferencias en la identidad personal tienen una significación de la que carecen las diferencias en la disposición temporal. La cuestión que hasta aquí he planteado es sólo esta: al explicar por qué el tiempo no puede tener significación racional, el teórico PI no puede utilizar el argumento obvio y mejor. No puede apelar al hecho de que un dolor no es menos doloroso porque esté menos próximo. Un dolor no es menos doloroso porque sea de otro.

El teórico PI podría decir:

> No entiendes mi argumento. El que un dolor esté más lejos en tu futuro no puede hacerlo menos doloroso *para ti*. Pero el que un dolor sea de otro lo hace menos doloroso *para ti*. Si es el dolor de otro, a ti no te hará ningún daño en absoluto.

La segunda de estas frases afirma dos extremos. El que un dolor esté más lejos en mi futuro no lo hace ni (*a*) menos doloroso en lo más mínimo, ni (*b*) menos mío en lo más mínimo. (*a*) es verdadera, pero irrelevante, puesto que la objeción a la que apela se aplica igualmente a la teoría del Propio Interés. El que un dolor sea de otro no lo hace en lo más mínimo menos doloroso. (*b*) también es verdadera. El hecho de que un dolor esté más lejos en mi futuro no lo hace menos *mi* dolor en absoluto. Pero esta verdad no es un argumento. Lo que el teórico PI necesita establecer, al atacar a Próximus, es que una diferencia en *quién* siente un dolor tiene gran significación racional, mientras que no puede haber significación racional en *cuándo* se siente un dolor. Todo lo que (*b*) subraya es que estas son diferencias *diferentes*. El tiempo no es lo mismo que la identidad personal. Por sí mismo, esto no puede mostrar que el tiempo es menos significativo.

Ahora resumiré estas afirmaciones. El teórico PI tiene que criticar a Próximus. Según PI, deberíamos tomar en consideración las diferencias tanto en la cualidad de doloroso como en la identidad

de los que sufren. Próximus también toma en cuenta diferencias en la disposición temporal. El teórico PI no ha mostrado que las diferencias en cuanto a la identidad personal tengan una significación racional de la que carezcan las diferencias en la disposición temporal. Puede haber argumentos a favor de esta tesis. Pero todavía no he aportado uno. El teórico PI no puede utilizar el mejor argumento. No puede descartar diferencias en la disposición temporal con la afirmación de que no son diferencias en la cualidad de doloroso. Ni tampoco diferencias en la identidad personal. Ni tampoco puede el teórico PI descartar diferencias en la disposición temporal con el fundamento de que no son diferencias en la identidad personal. El que estas sean diferencias *diferentes* no puede mostrar que una tenga una significación racional de la que la otra carece.

64. Sufrimiento pasado o futuro

El teórico PI podría sostener que no hay necesidad de ningún argumento. No podemos argumentarlo todo: algunas cosas tienen que ser dadas por supuestas. Y podría decir esto en relación con su tesis presente. Podría decir que, cuando comparamos las preguntas «¿A quién sucede?» y «¿Cuándo sucede?», vemos con claridad que sólo la primera tiene significación racional. Vemos con claridad que no es irracional preocuparse menos por un dolor si va a ser sentido por otra persona, pero que *es* irracional preocuparse menos meramente a causa de una diferencia en lo que respecta a *cuándo* un dolor es sentido por uno mismo.

¿Es así? La inclinación hacia lo próximo no es nuestra única inclinación respecto al tiempo. También estamos inclinados hacia el futuro. ¿Se trata de una actitud irracional?

Consideremos mis *Operaciones Pasada o Futura*.

> *Caso Uno*. Me encuentro en un hospital porque me van a someter a un cierto tipo de intervención quirúrgica. Como se trata de algo completamente seguro, y que siempre tiene éxito, no tengo ningún miedo al resultado. La intervención quirúrgica puede ser breve, o puede por el contrario llevar mucho tiempo. Como tengo que cooperar con el cirujano, no me pueden anestesiar. He pasado por esta intervención una vez, y puedo recordar lo dolorosa que es. Siguiendo un nuevo método, dado que la operación es tan dolorosa, a los pacientes ahora se les hace olvidarla después. Una droga les quita los recuerdos de las últimas horas.
>
> Me acabo de despertar. No puedo recordar haberme ido a dormir. Le pregunto a mi enfermera si ya han decidido cuándo va a ser mi operación y cuánto tiempo va a durar. Ella dice que conoce los hechos relativos a mí y a otro paciente, pero que no puede recordar qué hechos se aplican a quién. Sólo puede decirme que lo que sigue es verdadero. Puede que yo sea el paciente a quien operaron ayer. En ese caso, mi operación fue la más larga que jamás tuvo lugar, durando diez horas. En vez de eso puede que yo sea el paciente que va a sufrir una breve operación hoy dentro de un rato. Es verdadero que o bien yo sufrí durante diez horas o bien que sufriré durante una hora.
>
> Le pido a la enfermera que descubra cuál de las dos historias es verdadera. Mientras se va, se me hace claro cuál prefiero que sea verdadera. Si me entero de que la primera es verdadera, me sentiré muy aliviado.

Mi inclinación hacia el futuro me hace sentirme aliviado, entonces, al saber que mi dolor está en el pasado. Mi inclinación hacia lo próximo podría, del mismo modo, hacerme sentir aliviado por que un dolor se haya pospuesto. En cualquier caso, podría preferir una disposición temporal diferente para mi suplicio aunque, con la disposición temporal diferente, el suplicio fuera mucho peor. En comparación con una hora de dolor hoy dentro de un rato, yo podría, como Próximus, preferir diez horas de dolor el año que viene. O, como en el ejemplo, podría preferir diez horas de dolor ayer.

¿Es irracional esta segunda preferencia? ¿Debería por el contrario esperar ser el segundo paciente, cuyo dolor aún está por llegar? Antes de discutir esta cuestión, debería explicar un rasgo del caso: la amnesia inducida.

Hay autores que sostienen que, si alguna parte de mi futuro no va a estar enlazada por la memoria con el resto de mi vida, puedo ignorar racionalmente lo que me vaya a ocurrir durante ese período. Para estos autores, una dosis doble de amnesia es tan buena como

una anestesia. Si no tendré recuerdos mientras estoy sufriendo, y más tarde tampoco tendré recuerdos de mi sufrimiento, no tengo ninguna necesidad —mantienen— de preocuparme por este sufrimiento futuro. Esta es una tesis controvertida. Pero aunque esté justificada no se aplica a mi ejemplo. Este no implica una dosis *doble* de amnesia. Durante mi dolorosa operación conservaré todos mis recuerdos. Es cierto que después me harán olvidar la operación. Pero esto no elimina mi razón para estar preocupado por mi futuro sufrimiento. Si negáramos esto, tendríamos que afirmar que alguien no debería preocuparse cuando, sabiendo ya que está a punto de morir, se entera del hecho extra de que morirá dolorosamente. Más tarde no recordará *esos* dolores.

Si con la imaginación nos ponemos en el lugar del paciente que va a sufrir durante una hora hoy dentro de un rato, la mayoría de nosotros estaríamos preocupados. Estaríamos preocupados aunque supiésemos que más tarde no íbamos a recordar esta hora de dolor. Y ahora puedo explicar por qué mi ejemplo incluye la amnesia inducida. Esto nos da la comparación correcta. Si me he enterado de que soy el segundo paciente, me encuentro en el siguiente estado mental. Creo que tendré una hora de dolor hoy dentro de un rato y me puedo imaginar aproximadamente lo horrible que será el dolor. Esto basta para preocuparme. Si me he enterado en cambio de que soy el primer paciente, me encuentro en el estado mental estrictamente comparable. Creo que tuve diez horas de dolor ayer, y me puedo imaginar aproximadamente lo horrible que tuvo que haber sido el dolor. Mi estado mental difiere solamente en los dos respectos que estoy discutiendo. Mi creencia tiene un tiempo diferente, siendo sobre el pasado antes que sobre el futuro. Y es una creencia acerca de diez horas de dolor antes que acerca de sólo una. Haría confusa la comparación el que yo no creyera sólo que sufrí ayer, sino que pudiera también recordar el sufrimiento. Cuando pienso que sufriré hoy dentro de un rato, no tengo nada comparable a los recuerdos de este futuro sufrimiento. Y los recuerdos del dolor son muy diversos; algunos son en sí mismos dolorosos, otros no. Por eso libra al ejemplo de un rasgo irrelevante y que introduce complicaciones el que yo tenga sobre mi dolor pasado sólo lo que yo tengo

sobre mi dolor futuro: una creencia, con la capacidad de imaginar el carácter terrible del dolor.

La amnesia inducida purifica el caso. Pero todavía puede levantar suspicacias. Por eso añado

> *Caso Dos*. Cuando me despierto, recuerdo un prolongado período de sufrimiento que tuvo lugar ayer. Pero no puedo recordar lo que duró ese período. Le pregunto a la enfermera si mi operación está completada o si aún falta por hacer alguna intervención adicional. Como antes, ella conoce los hechos sobre dos pacientes, pero no puede recordar cuál soy yo. Si soy el primer paciente, tuve cinco horas de dolor ayer, y mi operación ha terminado. Si soy el segundo paciente, tuve dos horas de dolor ayer, y tendré otra hora de dolor hoy [33].

En el Caso Dos no hay amnesia; pero esto no supone ninguna diferencia. O sufrí durante cinco horas y no voy a tener que soportar más dolor, o sufrí durante dos horas y me queda otra hora de dolor. Una vez más preferiría que fuese verdadero lo primero. Preferiría que mi vida contuviese más horas de dolor si eso significa que no queda por venir más dolor como éste.

Si os pusierais con la imaginación en mi lugar en estos dos casos, la mayoría de vosotros tendría mi misma preferencia. Si no supiéramos si hemos sufrido durante varias horas, o si sufriremos dentro de un rato durante una hora, la mayoría de nosotros preferiría con mucho que lo primero fuese verdadero. Si pudiésemos hacerlo verdadero, indudablemente lo haríamos. Si fuéramos religiosos rezaríamos para que fuera verdadero. Según algunas explicaciones, esta es la única manera concebible de afectar al pasado. Dios puede haber hecho que un suceso pasado ocurra únicamente porque, en aquel momento, Él tenía conocimiento previo de nuestra posterior oración retrospectiva, y Él eligió conceder lo que esta oración imploraba. Aunque no creamos que nosotros podríamos hacer de esta manera, por la gracia de Dios, que nuestro dolor estuviera en el pasado, preferiríamos sin duda que estuviera en el pasado, incluso al coste de que durara diez veces más.

[33] Esta versión del caso me fue sugerida por G. Harman.

¿Es irracional esta preferencia? La mayoría de nosotros respondería No. Si acepta esta respuesta, el teórico PI tiene que abandonar su tesis de que la pregunta «¿Cuándo?» no tiene importancia racional. No puede mantener que una mera diferencia en la disposición temporal de un dolor, o en su relación con el momento presente, «no sea en sí misma un fundamento racional para tener más o menos consideración por él» [34]. Si un dolor está en el pasado o en el futuro es una mera diferencia en su relación con el momento presente. Y, si no es irracional preocuparse más de los dolores que están en el futuro, ¿por qué es irracional preocuparse más de los dolores que están en el futuro más próximo? Si el teórico PI admite como sostenible una desviación de la neutralidad temporal, ¿cómo puede criticar la otra?

65. La dirección de la causación

El teórico PI podría decir: «Como no podemos influir sobre el pasado, esta es una buena razón para estar menos preocupados por él. No hay una justificación así para la inclinación hacia lo próximo».

Esto puede contestarse. Para empezar, podemos señalar que seguimos predispuestos a favor del futuro aunque, como el pasado, no pueda modificarse. Supongamos que estamos en prisión, y que seremos torturados hoy dentro de un rato. En casos como este, cuando pensamos que nuestro sufrimiento futuro es inevitable, nuestra actitud hacia él no se compadece con nuestra actitud hacia el sufrimiento pasado. No pensaríamos, «Como la tortura es inevitable, eso equivale a que ya está en el pasado». Nos sentimos enormemente aliviados cuando esos dolores inevitables están en el pasado. En tales casos la inclinación hacia el futuro no puede justificarse apelando a la dirección de la causación. Nuestra razón para preocuparnos por esos dolores futuros no es que, a diferencia de los dolores pasados, podamos influir en ellos. Sabemos que *no podemos* influir en ellos. Nos preocupamos por esos dolores futuros simplemente porque todavía no están en el pasado.

[34] Rawls, p. 293.

El teórico PI podría replicar: «Tal justificación no es necesario que se mantenga en todos los casos. Cuando discutimos una actitud general, tenemos que contentarnos con una verdad general. Semejantes actitudes no pueden ser "ajustadas". Si los sucesos están en el futuro en *la mayoría* de los casos corresponde a si podemos o no podemos influir en ellos. Esto basta para justificar la inclinación hacia el futuro. Si carecemos de esta inclinación, estaríamos igual de preocupados por los dolores y los placeres pasados, a los que no podemos afectar. Esto distraería nuestra atención de los dolores y los placeres futuros, sobre los que sí podemos influir. Como estaríamos distraídos de este modo, tendríamos menos éxito en nuestros intentos de conseguir placeres futuros y evitar dolores futuros. Algo sería peor para nosotros».

Podríamos responder: «Si esto es verdadero, hay otra verdad similar. Si estuviéramos preocupados en la misma medida por los dolores y los placeres de nuestro futuro *más lejano*, esto distraería nuestra atención de los dolores y los placeres del futuro más próximo. Si queremos reducir nuestro sufrimiento futuro, deberemos prestar más atención a los posibles dolores del futuro más próximo, puesto que tenemos menos tiempo para evitar o reducir esos dolores. Una tesis similar se aplica a los placeres futuros. Nuestra necesidad de influir sobre el futuro más próximo es más urgente. Si tu tesis justifica la inclinación hacia el futuro, esta tesis justifica la inclinación hacia lo próximo».

Podríamos añadir: «Nos cuidamos más del futuro cercano incluso en los casos especiales en que no podemos influir en él. Pero estos casos corresponden a las circunstancias especiales en que no podemos afectar al futuro. Estas dos actitudes hacia el tiempo corresponden aproximadamente a estos hechos acerca de la causación. Tus tesis, por esta razón, no pueden demostrar que sólo se puede mantener una de estas actitudes».

El teórico PI podría decir: «Pasas por alto una diferencia. Nosotros podemos actuar directamente sobre la inclinación hacia lo próximo. Si está previsto que vayamos a tener una hora de dolor hoy dentro de un rato, tal vez podamos posponer ese dolor, al coste de hacerlo peor. Como Próximus, tal vez podamos cambiar ese

dolor por diez horas de dolor el año que viene. Pero no podemos cambiar este dolor por diez horas de dolor ayer. No podemos poner los dolores futuros en el pasado, al coste de hacerlos peores. La diferencia importante es esta. Como podemos afectar tanto al futuro próximo como al distante, nuestra inclinación hacia lo próximo a menudo nos hace obrar en contra de nuestros propios intereses. Esta inclinación es mala para nosotros. En contraste, como no podemos afectar al pasado, nuestra inclinación hacia el futuro nunca nos hace obrar en contra de nuestros intereses. Esta segunda inclinación no es mala para nosotros. Por eso sólo se puede mantener la segunda inclinación».

Frente a esto hay tres respuestas: (1) Este argumento tiene una premisa falsa. El hecho de que una actitud es mala para nosotros no demuestra que sea irracional. Puede demostrar como mucho que deberíamos tratar de cambiar esa actitud. Si asesinan a alguien a quien quiero, tal vez yo, al cabo de un tiempo, debería tratar de disminuir mi dolor. Pero esto no muestra que yo no tenga razón para afligirme. La pena no es irracional simplemente porque trae infelicidad. A la afirmación «Tu pesar es inútil», Hume contestó, «Muy cierto, justo por esa razón lo siento» [35].

(2) Aunque se niegue (1), este argumento falla. Asume que lo que importa es si algo es malo para nosotros. Esto da por bueno lo que hay que demostrar. El teórico PI condena la inclinación hacia lo próximo. Si tenemos esta inclinación, nos cuidamos más de nuestro futuro más cercano; y lo que será, en el conjunto, peor para nosotros puede ser mejor para nosotros en el futuro más cercano. Si nuestra inclinación es justificable, podemos entonces negar la asunción de que lo que importa es si algo será malo para nosotros. Como puede negarse esta asunción si nuestra inclinación es justificable, no puede ayudarnos a demostrar que nuestra inclinación *no* sea justificable.

(3) No se ha demostrado que la inclinación hacia lo próximo *sea* mala para nosotros. Puesto que tenemos una necesidad más urgente de influir sobre el futuro más próximo, la inclinación hacia lo próximo es de alguna manera buena para nosotros. Pero supongamos que esta inclinación es, una vez considerados todos los factores, mala para nosotros. *También lo es la inclinación hacia el futuro*. Como explicaré después, sería mejor para nosotros si no nos preocupáramos más por el futuro. El argumento anterior tiene otra premisa falsa. No es cierto que la inclinación hacia el futuro no sea mala para nosotros.

[35] Hume (2), p. 177.

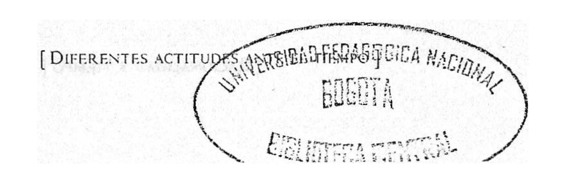
El teórico PI tiene que condenar la inclinación hacia lo próximo. Si su crítica apela a la neutralidad temporal, también tiene que criticar la inclinación hacia el futuro. Apelando a hechos acerca de la causación, el teórico PI trató de evitar esta conclusión. Pero este intento ha fracasado.

Al condenar la inclinación hacia lo próximo, el teórico PI podría decir: «Como nuestra necesidad de afectar a lo próximo es más urgente, la inclinación hacia lo próximo es muy natural. No sorprende que la Evolución dé esta inclinación a todos los animales. Pero, como nosotros somos racionales, podemos levantar la cabeza sobre lo que heredamos de la Evolución, y revisarlo críticamente. Podemos ver que esta inclinación no puede ser racional. El que un dolor esté en el futuro más cercano no puede ser una *razón* para preocuparnos más de él. Una mera diferencia en disposición temporal no puede tener significación racional» [36].

Si el teórico PI hace esta afirmación, tiene que hacer una afirmación similar acerca de la inclinación hacia el futuro. Podría decir: «Como no podemos afectar al pasado, es natural que nos preocupemos menos por él. Pero esta inclinación no puede ser racional. Esto resulta de lo más claro cuando no podemos afectar al futuro. El que un dolor inevitable esté en el futuro más bien que en el pasado no puede ser una *razón* para preocuparse más por él. Es irracional sentirse aliviado cuando está en el pasado».

En Mis Operaciones Pasada y Futura, yo prefería que fuese cierto que sufrí durante varias horas ayer, en vez de que lo fuese que voy a sufrir una hora hoy dentro de un rato. Esta no es una prefe-

[36] Repito algunas observaciones de R. Nozick.



rencia sobre la base de la cual yo pudiera actuar. Pero esto es irrelevante en relación con la cuestión de si esta preferencia es irracional. El teórico PI no puede afirmar que *no* es irracional *porque* no puedo actuar a partir de ella. Él podría decir, «¡Qué preferencia más absurda! Deberías estar agradecido por no poder actuar a partir de ella». Y esto es lo que él *tiene* que decir, si mantiene su afirmación de que nuestro interés por nosotros mismos debería ser temporalmente neutral. Si condena la inclinación hacia lo próximo porque no puede tener significación racional *cuándo* se siente un dolor, tiene que condenar la inclinación hacia el futuro. Tiene que establecer que es irracional sentirse aliviado cuando un dolor está en el pasado. La mayoría de nosotros encontraría esto difícil de creer. Si el teórico PI insiste en que deberíamos ser temporalmente neutrales, la mayoría de nosotros estaría en desacuerdo.

66. Neutralidad temporal

El teórico PI podría cambiar su posición. Podría condenar la inclinación hacia lo próximo, no a partir de la razón general de que la pregunta «¿cuándo?» no puede tener significación racional, sino a partir de una razón más particular.

Podría pasarse al otro extremo, y afirmar que la neutralidad temporal es inconcebible. Podría afirmar que es inconcebible que carezcamos de la inclinación hacia el futuro. Si esto fuese cierto, él podría criticar de nuevo sólo una de estas dos actitudes. No puede ser irracional tener una actitud si no es concebible que carezcamos de esa actitud. Pero, a diferencia de la inclinación hacia el futuro, la inclinación hacia lo próximo es claramente algo de lo que podríamos carecer. Podríamos estar interesados por igual en todas las partes de nuestro futuro. Algunas personas lo están. El teórico PI podría afirmar que este es el único patrón racional de intereses.

¿Es concebible que pudiéramos carecer de la inclinación hacia el futuro? Nuestras actitudes respecto del pasado no podrían ser iguales que nuestras actitudes ante el futuro. Algunas emociones o reacciones presuponen creencias sobre la causación. Como no podemos

afectar al pasado, estas emociones y reacciones no podrían mirar hacia atrás. De modo que no podríamos formarnos la intención de haber hecho algo ayer, o decidirnos firmemente a mejorar lo que ha quedado detrás de nosotros.

¿Hay estados mentales que miran esencialmente hacia delante, de un modo que no puede ser explicado por la dirección de la causación? Esta es una pregunta vasta, a la que no tengo necesidad de dar una respuesta completa. Bastará con que consideremos los estados mentales más importantes que están involucrados en nuestra inclinación hacia el futuro.

Uno de estos es el deseo. Hay giros de nuestro lenguaje que sugieren que los deseos miran esencialmente hacia delante. Comparemos «Quiero ir a Venecia el próximo invierno» con «Quiero haber ido a Venecia el último invierno». La segunda afirmación es oscura.

Nuestro lenguaje es aquí engañoso. Consideremos

> *Mi Deseo Temporalmente Neutral.* Me entero de que una vieja amiga se está muriendo en un país lejano. Nos habíamos separado enfadados, cosa de la que ahora me culpo. Tras enterarme de que mi amiga se está muriendo, surge en mí el imperioso deseo de pedirle que me perdone. Como no puedo comunicarme con ella por teléfono, lo mejor que puedo hacer es mandarle una carta urgente, pidiéndole que me perdone y despidiéndome de ella. Una semana después, no sé si mi amiga está todavía viva o ha recibido mi carta. Mi deseo más imperioso es que reciba mi carta antes de morir.

Si los deseos miran esencialmente adelante, tengo según todos los indicios que estar en dos estados mentales: un deseo condicional, y una esperanza condicional. Hay que decir que tengo que querer que, en caso de que esté con vida, mi amiga reciba la carta antes de morir, y tengo que esperar, en caso de que haya muerto, que haya recibido la carta antes de morir. Pero esta descripción, aunque venga requerida lingüísticamente, es engañosa. Distinguir aquí dos estamos mentales, el deseo y la esperanza, significa subdividir lo que en su naturaleza es un estado singular. Mi «esperanza» es en su naturaleza y en su fuerza justo como mi «deseo». Lo que quiero es que

a mi amiga le llegue la carta antes de morir. Con tal que estos sucesos ocurran en este orden, me da lo mismo que estén en el pasado o en el futuro.

Aunque suponga cambiar el concepto, es por tanto mejor decir que podemos tener deseos sobre el pasado. Y no está claro que esto sea un cambio. Por ejemplo, puedo querer que sea verdadero que, en mi borrachera de la pasada noche, no me haya portado de forma vergonzosa. Y puedo querer que esto sea cierto por sí mismo, no a causa de sus posibles efectos en mi futuro. De forma similar, tras leer las cartas de Keats o de van Gogh, puedo querer que sea cierto que supieran lo grandes que fueron sus logros.

En estos ejemplos, no conozco la verdad. Supongamos que supiera que me porté de forma vergonzosa la noche pasada. ¿Puedo querer que sea verdad lo contrario? Sería más natural llamar a esto un *deseo irrealizable*. Pero esta distinción tampoco parece importante. Cuando me entero de que me porté de forma vergonzosa, mi deseo de que no hubiera sido así se convierte en un deseo irrealizable. Pero el deseo irrealizable puede no ser más débil que el deseo corriente.

Puede cambiar el concepto de deseo el que afirmemos que podemos querer que algo que sabemos que es falso sea verdadero. No necesitamos decidir si este cambio en nuestro concepto sería una mejora. Estoy discutiendo la cuestión diferente de si podemos tener deseos acerca del pasado. He afirmado que podemos, aunque algunos giros de nuestro lenguaje sugieran que no. Podemos expresar tales deseos con otros giros de nuestro lenguaje. Podemos decir, como he hecho yo, que queremos que sea verdadero que algún suceso ocurriera o no ocurriera.

Puede objetarse que los deseos están esencialmente vinculados a actos posibles. Esto es como la tesis de que «deber» implica «poder». Según esta opinión, no podemos tener deseos a partir de los que sea imposible obrar. De esta tesis general podríamos deducir la tesis especial de que no podemos tener deseos sobre el pasado, puesto que no podemos afectar al pasado [37].

[37] Una versión más sutil de esta objeción se presenta en Edidin. Yo no respondo del todo a la objeción de Edidin.

Esta tesis general es falsa. Desde luego que hay estrechas conexiones entre deseos y actos. Si queremos intensamente que algo se cumpla, trataremos de descubrir si podemos hacerlo verdadero nosotros. Y «el signo primitivo de querer es tratar de conseguir» [38]. Pero aquí el deseo viene primero. No tenemos que saber si podríamos realizar algo antes de que podamos querer que se realice.

Podemos admitir una manera en que los deseos están vinculados a actos. Si las personas no pudieran actuar no podrían tener deseos. No podríamos tener el concepto de deseo a no ser que tuviéramos también el concepto de acto. Pero podemos tener un deseo *particular* sin poder actuar a partir de él. Podemos querer que algo se realice aunque sepamos que ni nosotros ni nadie podría posiblemente haberlo realizado. Los pitagóricos deseaban que la raíz cuadrada de dos fuese un número racional. Es lógicamente imposible que tal deseo se realice. Como podemos tener deseos que ni siquiera un Dios omnipotente podría realizar, los deseos particulares no están vinculados a actos posibles. Esto elimina esta razón para negar que podamos tener deseos sobre el pasado.

A continuación podemos considerar los estados mentales que son los más importantes en esta discusión: la espera ilusionada de un suceso futuro, y su contrapartida negativa, la anticipación penosa o angustiosa. Estos dos estados mentales están esencialmente dirigidos al futuro. Pero esto puede que sea otra verdad superficial. ¿Podría haber estados comparables dirigidos hacia el pasado?

Puede pensarse que en realidad sí que tenemos tales estados que miran atrás. La inclinación hacia el futuro no se aplica a muchos tipos de sucesos, como los que nos producen orgullo o nos dan vergüenza. Pero aunque el conocimiento de un logro pasado puede darnos placer, esto no es análogo a la espera ilusionada. Estamos discutiendo nuestra actitud, no ante el *hecho* de que nuestra vidas contengan ciertos tipos de sucesos, sino ante nuestra *experiencia* en otros momentos de vivir estos sucesos. En aras de la simplicidad, he estado discutiendo las actitudes que tenemos ante las experiencias

[38] Anscombe (1), p. 68.

que son meramente en sí mismas agradables o penosas. ¿Contemplamos en efecto retrospectivamente los placeres pasados del mismo modo en que esperamos con impaciencia los futuros placeres?

Una vez más, los recuerdos hacen surgir una complicación. Estos pueden ser en sí mismos agradables o penosos. Podemos disfrutar al recordar placeres y contrariarnos al recordar dolores. Pero ninguno de estos es estrictamente análogo a los dolores y los placeres de la anticipación. Por eso necesitamos considerar nuestra actitud ante los dolores y placeres pasados que conocemos, pero de los que no tenemos recuerdos penosos o agradables.

Consideremos *Mis Pasados Suplicios*

> *Caso Uno*. Soy muy olvidadizo. Me preguntan, «¿Puedes recordar lo que te ocurrió el mes de mayo de hace diez años?». Me doy cuenta de que no puedo recordar nada sobre ese mes. Entonces me dicen que, al comienzo de ese mes, descubrieron que yo tenía una enfermedad que requería cuatro semanas de tratamiento inmediato, un tratamiento muy doloroso. Como tuvo un éxito total, no tengo razón alguna para tener miedo del futuro. Cuando me recuerdan este hecho, surge en mí un recuerdo tenue, que no es en sí penoso.

Se me ha recordado, para mi sorpresa, que hace diez años pasé un mes de horrible dolor. Todo lo que me queda ahora es un tenue recuerdo de este hecho, y la habilidad de imaginar lo malo que tuvo que haber sido mi dolor. Cuando me recuerdan este suplicio pasado, ¿me siento afectado? ¿Tengo acaso lo que correspondería a la anticipación penosa? No. Reacciono a este recuerdo con *completa indiferencia*.

Si me enterara de que dentro de diez años pasaré un mes de sufrimiento atroz, *no* me quedaría indiferente. Me angustiaría. Pero no me angustiaría en absoluto si se me recordara que hace diez años pasé un mes así.

Como estamos inclinados hacia lo próximo, puede ser útil que consideremos el

> *Caso Dos*. Me despierto un día que creo que es el Primero de Mayo. Pero en realidad es el Primero de Junio. Acabo de pasar un mes similar de un tratamiento muy doloroso pero con un éxito total. Para que no tenga recuerdos penosos, me han hecho olvidar este mes entero.

Me entero de que acabo de pasar un mes de sufrimiento atroz. Aquí tampoco consideraría esto como una mala noticia. Más exactamente, lamentaría el hecho de que un mes de mi vida hubiera tenido que gastarse de ese modo. Podría estar en cierto modo preocupado por el supuesto éxito de este tratamiento. Y podría tener miedo de que, si la amnesia inducida no dura, más tarde tendré recuerdos penosos de este tratamiento. Pero no estaría angustiado en absoluto por el hecho de que, durante este mes, sufrí atrozmente. Consideraría este sufrimiento reciente con completa indiferencia. En contraste, si me enterara de que estaba a punto de pasar por una experiencia así de horrorosa, me angustiaría mucho.

Puede ser una objeción al Caso Dos el que conlleve amnesia inducida. Por eso añado el

> *Caso Tres*. En mi vida real, a menudo he sufrido fuertes dolores. Puedo recordar estos dolores, pero estos recuerdos no son en sí mismos dolorosos. El peor sufrimiento que puedo recordar duró tres días en 1979.

Cuando me recuerdo a mí mismo estos tres días tan dolorosos, no me aflijo en absoluto. En los Casos imaginarios Uno y Dos, creo que consideraría con indiferencia mis suplicios pasados. En mi vida real, considero en efecto mi sufrimiento pasado con una indiferencia absoluta.

Creo que, en este aspecto, la mayoría de la gente es como yo. A no ser que sus recuerdos sean dolorosos, las personas consideran sus sufrimientos pasados con indiferencia. Conozco a pocas personas cuya reacción sea diferente. Estas personas afirman que, aunque no tengan recuerdos dolorosos, encuentran el conocimiento de sus dolores pasados moderadamente angustioso. Pero no sé de nadie que tenga lo que correspondería del todo a los sufrimientos de la anticipación.

Efectivamente, no tenemos esta actitud ante nuestros dolores pasados. Y no recordamos los placeres del modo en que los esperamos. ¿Podrían darse tales estados mentales? ¿Podría ser el «mirar hacia atrás» un suceso pasado, quitando su dirección temporal, como verlo en el futuro?

Podríamos decir: «Esperamos con ansia un suceso futuro cuando el pensar en él nos da placer. El pensar en un suceso pasado nos podría dar un placer similar. Y a los sufrimientos de la anticipación les podrían corresponder dolores de retrospección.

Puede objetarse: «Subestimas lo que conlleva esperar con ansia. No es meramente verdadero que el pensamiento de placeres futuros nos dé placer. *Anticipamos* estos placeres. De forma similar, anticipamos dolores. La *anticipación* no puede tener un equivalente retrospectivo».

Podríamos responder: «Tal vez seamos incapaces de imaginar cómo sería tener este equivalente. Pero esto no demuestra que no se pudiera tener. Los ciegos congénitos no pueden imaginar cómo es ver».

Puede que esta réplica no elimine la objeción. Si es así, nuestras tesis pueden revisarse. Aunque mirar hacia atrás no sea *justo como* mirar hacia delante, podría ser igualmente bueno, o, en el caso de los dolores, igualmente angustioso. Esto llevaría consigo un cambio en nuestras actitudes. Pero este cambio *es* concebible. Podemos describir con claridad a alguien que, en este aspecto, es diferente de nosotros. Cuando a una persona así se le recuerda que una vez tuvo un mes de sufrimiento, se angustia tanto como cuando se entera de que más adelante tendrá un mes de sufrimiento. Es neutral, de forma similar, respecto de los sucesos agradables. Cuando se le dice que más adelante tendrá un período de gran goce, le agrada enterarse de esto. Espera con gran impaciencia este período. Cuando se le recuerda que una vez tuvo este período, le agrada igualmente. Llamaré a este hombre imaginario *Sintiempo*.

Este hombre es muy diferente de nosotros. Pero su descripción es coherente. Podemos por tanto rechazar la sugerencia que se presentó arriba. Es concebible que pudiésemos carecer de la predisposición hacia el futuro. Aunque no pudiésemos ser temporalmente neutrales por completo, podríamos haber sido como Sintiempo.

67. Por qué no deberíamos estar predispuestos a favor del futuro

Nuestra predisposición a favor del futuro es mala para nosotros. Sería mejor para nosotros que fuéramos como Sintiempo. En algunas cosas saldríamos perdiendo. Por ejemplo, no nos sentiríamos aliviados cuando las cosas malas quedaran en el pasado. Pero también saldríamos ganando. No nos pondríamos tristes cuando las cosas buenas quedaran en el pasado.

Las ganancias compensarían las pérdidas. Una razón sería esta. Cuando miramos atrás, nos podemos permitir ser selectivos. Debemos recordar algunos de los sucesos malos de nuestra vida cuando esto nos pueda ayudar a evitar las repeticiones. Pero podemos permitirnos el lujo de olvidar la mayoría de las cosas malas que han ocurrido, mientras que conservamos por repetición todos los recuerdos de las cosas buenas. Sería malo para nosotros que fuéramos tan selectivos cuando miramos al futuro. Como no pensemos en todas las cosas malas que en principio pueden suceder, perderemos nuestra oportunidad de prevenirlas. Puesto que no debemos ser selectivos cuando miramos al futuro, pero podemos permitirnos serlo cuando miramos al pasado, esto último sería, en general, más agradable [39].

Habría otras ganancias mayores. Una estaría en nuestra actitud ante el envejecimiento y la muerte. Vamos a considerar para empezar el argumento con el que Epicuro estableció que nuestra futura inexistencia no puede ser algo que lamentemos. Nosotros no lamentamos nuestra inexistencia pasada. Por tanto, ¿por qué deberíamos lamentar nuestra inexistencia futura? Si consideramos a la una con ecuanimidad, ¿por qué no deberíamos hacer extensible esta actitud a la otra?

[39] Creo que estas ganancias compensarían por sí mismas las pérdidas descritas por el teórico PI al comienzo de la Sección 65.

Hay quienes mantienen que este argumento falla porque, mientras que podríamos vivir más, no podríamos haber nacido antes. Pero esta no es una buena objeción. Cuando se dieron cuenta de que la raíz cuadrada de dos no era un número racional, los pitagóricos lo lamentaron. Podemos lamentarnos de las verdades, aunque sea lógicamente imposible que estas verdades sean falsas.

El argumento de Epicuro falla por una razón diferente: estamos inclinados a favor del futuro. Puesto que tenemos esta predisposición, el mero conocimiento de que una vez sufrimos puede que ahora no nos altere. Pero nuestra ecuanimidad no demuestra que nuestro sufrimiento pasado no fuera malo. Lo mismo podría ser verdad de nuestra pasada inexistencia. El argumento de Epicuro, por tanto, sólo tiene fuerza para las personas que carecen de la predisposición a favor del futuro y no lamentan su inexistencia pasada. Como esas personas no existen, el argumento no tiene fuerza para nadie.

Aunque el argumento falla, puede proporcionar cierto consuelo. Si nos da miedo la muerte, el argumento demuestra que el objeto de nuestro pavor no es *nuestra inexistencia*. Es sólo nuestra *futura* inexistencia. El que podamos pensar con serenidad en nuestra pasada inexistencia no demuestra que no sea algo lamentable. Pero como, en efecto, no contemplamos con pavor nuestra inexistencia pasada, tal vez podamos usar este hecho para reducir nuestro terror, o depresión, cuando pensamos sobre nuestra muerte inevitable. Si a menudo pensamos, contemplándola serenamente, sobre la oscuridad detrás de nosotros, algo de esta serenidad puede transferirse a nuestra contemplación de la oscuridad ante nosotros.

Supongamos ahora que carecemos de la predisposición a favor del futuro. Somos como Sintiempo. Entonces saldríamos ganando enormemente en nuestra actitud ante el envejecimiento y la muerte. A medida que nuestra vida pasa, tendríamos cada vez menos que anticipar del futuro y cada vez más que contemplar en el pasado. Este efecto será más claro si imaginamos otra diferencia. Supongamos que nuestras vidas comenzaran, no con el nacimiento y la niñez, sino como comenzó la de Adán. Supongamos que, aunque seamos adultos y tengamos el conocimiento y las destrezas de los

adultos, hemos empezado a existir ahora mismo. Carecemos de predisposición a favor del futuro. ¿Acaso nos debería alterar mucho el pensamiento de que ayer no existíamos?

Esto depende de lo que va mal con la inexistencia. Algunos piensan que es mala en sí misma. Pero la opinión más plausible es que su único defecto es las cosas que nos hace perder. Supongamos que aceptamos esta opinión. Tal vez pensemos entonces que es una razón para lamentarnos el que nuestra vida sea finita, limitada en ambos extremos por la inexistencia. Pero si hubiéramos empezado a existir ahora mismo, no pensaríamos que hay algo malo tras nosotros. Nuestra razón para lamentarnos sería simplemente que nos habíamos perdido muchas cosas que hubieran sido buenas. Supongamos que yo ahora pudiera ser como soy en realidad aunque hubiera venido al mundo como uno de los pocos privilegiados en torno a 1700. Entonces lamentaría enormemente haber nacido realmente en 1942. Preferiría con mucho haber vivido los dos siglos y medio anteriores, habiendo contando entre mis amigos a personas como Hume, Byron, Chejov, Nietzsche y Sidgwick.

En mi caso imaginario, no estamos predispuestos a favor del futuro ni hemos comenzado a existir ahora mismo. Aunque lamentaríamos el hecho de que no hubiéramos existido antes, no nos molestaría mucho el pensamiento de que tan sólo ayer no existíamos. No consideraríamos este hecho con la clase de pavor o pesar con el que la mayoría de las personas reales considerarían la inesperada perspectiva de su muerte al día siguiente. No tendríamos ese pavor o ese pesar porque, aunque no tendríamos nada bueno que recordar, tendríamos nuestra vida entera para anticipar.

Supongamos ahora que nuestras vidas casi se han acabado. Moriremos mañana. Si no estuviéramos predispuestos a favor del futuro, nuestra reacción reflejaría la que acabo de describir. No nos incomodaría mucho el pensamiento de que pronto dejaremos de existir, porque aunque ahora no tenemos nada que anticipar, tenemos nuestra vida entera para recordar.

Puede objetarse: «Puedes recordar ahora. Pero cuando estés muerto no podrás recordar. Y estarás muerto mañana. De manera que debes estar muy alterado». Podríamos responder: «¿Por qué? Es

cierto que cuando hayamos dejado de existir nunca podremos disfrutar recordando nuestra vida. Ahora no tenemos nada en absoluto que esperar del futuro, ni siquiera los placeres de mirar atrás. Pero era igualmente cierto que, antes de que empezáramos a existir, no podíamos disfrutar anticipando nuestra vida. Justo después de que empezásemos a existir, no teníamos nada en absoluto que recordar, ni siquiera los placeres de la anticipación. Pero entonces eso no fue razón para estar muy alterado, puesto que entonces podíamos anticipar nuestra vida entera. Como ahora podemos recordar nuestra vida entera, ¿por qué debería el hecho paralelo —de que no tenemos nada que esperar— darnos razón para estar muy alterados?».

Este razonamiento ignora las emociones que están esencialmente dirigidas al futuro. No se aplicaría a las personas para las que la alegría de mirar hacia el futuro viene de hacer planes, o de saborear las alternativas. Pero el razonamiento parece ser correcto cuando se aplica a personalidades más pasivas, esas que toman los placeres de la vida como vienen. Y, como esto es en parte cierto de nosotros, este razonamiento demuestra que seríamos más felices si no tuviéramos la predisposición a favor del futuro. Nos deprimiría mucho menos el envejecer y la proximidad de la muerte. Si fuéramos como Sintiempo, estar al final de nuestra vida sería más como estar al principio. En cualquier punto de nuestra vida podríamos disfrutar mirando o bien al pasado o bien al futuro en el todo de nuestra vida.

He establecido que, si careciéramos de la predisposición a favor del futuro, sería mejor para nosotros. Esto se ajusta a la tesis plausible de que sería mejor para nosotros que no tuviéramos la predisposición a favor de lo próximo. No hay razón aquí para criticar sólo la última predisposición. Las dos actitudes ante el tiempo son, en general, malas para nosotros.

Como creo que esta actitud es mala para nosotros, creo que no debemos estar predispuestos a favor del futuro. Esta creencia no da nada por sentado en relación con la racionalidad de esta predisposición. Según cualquier concepción moral plausible, sería mejor si fuéramos más felices. Este es el sentido en que, si pudiéramos, no

deberíamos estar predispuestos a favor del futuro. Al darnos esta predisposición, la Evolución nos niega la mejor actitud ante la muerte.

68. El paso del tiempo

Volvamos a mi cuestión principal. ¿Son irracionales estas actitudes ante el tiempo? La mayoría de nosotros piensa que la predisposición a favor del futuro no es irracional. Estamos inclinados a creer que sería irracional *carecer* de ella. Así que puede que no nos haya convencido para nada el razonamiento que di en el caso recién imaginado, en que somos temporalmente neutrales y moriremos mañana. Podemos describir a alguien a quien no le importa mucho la perspectiva de morir al día siguiente, porque ahora puede recordar su vida entera. Pero esta actitud, aunque describible, puede parecer loca, o llevar consigo un error absurdo.

Nos será de utilidad construir un caso más simple, que no implica inexistencia ni nuestras actitudes ante una vida entera. Puede ser una variante de un ejemplo anterior, que se refiere a nuestro hombre imaginario que es temporalmente neutral. Consideremos

> *Cómo Recibe Sintiempo las Buenas Noticias.* Sintiempo está en el hospital para someterse a una dolorosa operación, que irá seguida por una amnesia inducida. Cierto día se despierta, sin ningún recuerdo en especial del día anterior. Le pregunta a la enfermera cuándo tendrá que sufrir esa dolorosa operación, y cuánto tiempo durará. Como antes, la enfermera conoce los hechos acerca de dos pacientes, pero no está segura de cuál de los dos es él. En cualquier caso, sin embargo, la operación tenía que ser excepcionalmente larga, durando diez horas completas. Y la enfermera sabe que una de las dos cosas siguientes es verdadera: o bien él sufrió ayer durante diez horas, o bien él sufrirá hoy dentro de un rato durante diez horas.
>
> Sintiempo se hunde en la tristeza. Había esperado una operación más corta.

Cuando vuelve la enfermera exclama «¡Buenas noticias! Tú eres el que sufrió ayer».

> Sintiempo sigue igual de triste. «¿Por qué es eso una buena noticia?», pregunta. «Mi suplicio es igual de doloroso e igual de largo. Y forma parte de mi vida exactamente igual. ¿Por qué iba a suponer una diferencia para mí ahora que mi suplicio esté en el pasado?».

La amnesia inducida puede ser una objeción a este caso. Así que añado el

> *Caso Dos.* Sintiempo se somete a esta operación, y no tiene amnesia. Le visitamos el día antes del suplicio, y el día después. El día después, Sintiempo está igual de apesadumbrado. «¿Por qué debería sentirme aliviado?», pregunta. «¿Por qué es mejor que mi suplicio esté en el pasado?».

¿Está Sintiempo cometiendo un error? ¿Debe sentirse aliviado? La mayoría de nosotros contestaría Sí. Pero es difícil explicar por qué, sin dar por supuesto precisamente lo que hay que explicar. Podríamos decir, «Si el suplicio estuviera en su futuro, todavía tendría que sufrirlo. Como está en el pasado, se ha terminado, está acabado». Esta no es una explicación adicional de por qué Sintiempo es irracional. Qué él «todavía» tenga que sufrir el dolor se limita a repetir que el dolor está en su futuro.

Podríamos apelar aquí a lo que se llama el *paso del tiempo*, o la *objetividad del devenir temporal*. Podríamos decir: «Si su dolor está en el futuro, se aproximará más y más, hasta que él lo acabe sufriendo efectivamente. Pero, si su dolor está en el pasado, únicamente se alejará cada vez más». Estas observaciones parecen expresar una profunda verdad. Pero se trata de una verdad curiosamente escurridiza. ¿Qué se quiere decir con la frase «se aproxima más y más»? ¿No significa esto simplemente que, en momentos futuros, el futuro dolor estará más cerca de lo que entonces será el momento presente? Pero en momentos pasados un dolor pasado estuvo más cerca de lo que fue entonces el momento presente. ¿Dónde está la asimetría?

Es natural, como respuesta, utilizar cierta metáfora: la del movimiento a través del tiempo. Podríamos decir que nos estamos moviendo a través del tiempo dentro del futuro, o que los sucesos futuros se están moviendo a través del tiempo dentro del presente, o que la condición de presente, el alcance del «ahora», se mueve en el futuro. «Ahora» corre la secuencia de sucesos históricos, «como el foco luminoso que va bajando por la fila de coristas».

Puede servir la comparación de «ahora» con «aquí». Para los que niegan el paso del tiempo, o la objetividad del devenir temporal, «aquí» y «ahora» son estrictamente análogos. Ambos son relativos a los pensamientos, o a las proferencias, de un pensador particular. «Aquí» refiere al lugar en que está este pensador en un momento determinado, y «ahora» refiere al momento en que un pensamiento particular, uno que implique el concepto «ahora», es pensado. Las dos palabras podrían ser remplazadas por «este», como en la jerga del locutor «en este lugar y en este momento» [40].

Los que creen en el paso del tiempo rechazarían esta analogía. Admitirían que, en un universo que no contuviese pensadores, el concepto «aquí» carecería de aplicación. Pero sostienen que, incluso en un universo tal, todavía sería verdadero que ciertas cosas están ocurriendo *ahora,* y entonces sería verdadero que otras cosas están ocurriendo *ahora,* y entonces sería verdadero que otras cosas están ocurriendo *ahora,* y así sucesivamente. Aun en un universo sin vida, el alcance de «ahora» se movería todavía a través del tiempo desde el pasado al futuro.

La metáfora del movimiento a través del tiempo bien puede ser insostenible. ¿A qué rapidez nos movemos a través del tiempo? Tal vez no nos quedemos satisfechos con la única respuesta posible, «A un promedio de un segundo por segundo». Podemos afirmar que, *si* o bien nosotros o bien «ahora» nos podemos mover a través del tiempo, tiene que tener sentido el que este movimiento sea más rápido o más lento, pero esto no tiene sentido.

[40] Estas observaciones simplifican demasiado las cosas. Aunque neguemos el transcurso del tiempo, podemos admitir que «ahora» puede aplicarse a descripciones de un universo sin vida. El Suceso *X* es «ahora» relativo al Suceso *Y* si los dos ocurren al mismo tiempo.

Puede que estén justificados los críticos de la metáfora. Pero quizás esto no demuestre que no haya tal cosa como el paso del tiempo, o la objetividad del devenir temporal. Quizás sea esta una verdad categórica, en un nivel tan profundo que no deberíamos esperar que pudiera ser explicada, o bien por metáforas o bien en otros términos [41].

No trataré de decidir dónde está la verdad en este debate. Tomaré en consideración, por ello, ambas alternativas. Supongamos en primer lugar que, como piensan algunos filósofos y físicos, el paso del tiempo es una ilusión. Si esto es cierto, la neutralidad temporal no puede ser irracional. Al defender la teoría del Propio Interés, el teórico PI tiene que condenar la inclinación hacia lo próximo. Si la neutralidad temporal no puede ser irracional, el teórico PI podría volver a su opinión anterior de que tal neutralidad está racionalmente requerida. Tiene entonces que defender que, del mismo modo que es irracional sentirse aliviado cuando un dolor inevitable ha sido pospuesto, es irracional sentirse aliviado cuando está en el pasado. Encontraremos esto difícil de creer.

Supongamos a continuación que tenemos razón en creer en el paso del tiempo, o en la objetividad del devenir temporal. El teórico PI podría entonces retener su opinión posterior y apelar al paso del tiempo. Todavía tiene que condenar la inclinación hacia lo próximo. Podría afirmar: «Mientras que tienes excelentes razones para preocuparte menos de los dolores de los demás, no puedes, de un modo racional, preocuparte menos por dolores que son tuyos pero que se hallan distantes en el futuro. La mera distancia respecto del momento presente no puede tener significación racional». El teórico PI podría apoyar ahora esta tesis de una manera diferente. Podría abandonar la apelación a la neutralidad temporal —la tesis de que la mera disposición temporal no puede tener significación racional—. En vez de ello podría discriminar entre diferentes tipos de relación temporal.

Deberíamos recordar aquí que la mayoría de nosotros tiene una tercera actitud ante el tiempo: la predisposición a favor del presente. Si la mera disposición temporal no puede tener significación racional, no puede ser racional preocuparse más de los dolores presentes. El que yo esté *ahora* sufriendo un dolor horroroso no puede ser una razón para estar más preocupado ahora por este dolor horroroso. Esto puede parecer absurdo. El requisito de neutralidad temporal puede que parezca muy poco plausible cuando lo aplicamos a la predisposición a favor del presente. ¿Cómo puede ser irracional que me importe más mi horrible dolor mientras lo estoy sufriendo? Semejante tesis parece socavar toda la estructura del interés. El dolor importa sólo a causa de cómo nos sentimos cuando tenemos *ahora* dolor. Nos preocupamos de dolores futuros sólo porque, en el futuro, serán dolores *presentes*. Si los dolores futuros se comportaran como la *Mermelada Mañana* de Alicia, permaneciendo perpetuamente futuros, no importarían en absoluto [42].

[41] Cf. Pears (1), p. 249:

«Y, ya que el tiempo es una categoría que figura de una forma que no resulta incómoda en toda nuestra experiencia, es posible definir el pasado y el futuro de muchos modos diferentes, pero ninguna de estas definiciones es muy convincente. Son... por decirlo de alguna forma, débiles tautologías. Y son débiles tautologías no sólo porque estamos tan acostumbrados a emplear palabras temporales correctamente que no necesitamos recordatorios fuertes, sino también porque su estructura es especial. Y es que la mayor parte de las tautologías están construidas como columnas, por el procedimiento de colocar unos términos directamente encima de los otros como tambores de mármol. Pero las tautologías que dan la lógica de las palabras temporales juntan sus términos como las piedras de una bóveda. Ninguna conjunción singular de términos es indispensable ni podría sostenerse sola. Pero juntas forman el techo abovedado sobre el que está pintado el fresco del conocimiento».

Para discusiones sobre el transcurso del tiempo, véase Gale (2) (sobre todo los artículos de Williams y Grunbaum), Gale (1), y Smart (2).

[42] Estas observaciones son demasiado burdas. Aunque en el momento presente una experiencia sea neutra, o incluso desagradable, puede ser recordada después con gran deleite. Proust (1), p. 97, escribe:

«... en una fecha muy posterior, cuando repasé gradualmente, en orden inverso, los momentos por los que yo había pasado antes de estar tan enamorado de Albertine, cuando mi corazón lleno de cicatrices podía distanciarse él mismo sin sufrimiento de Albertine muerta, entonces fui yo capaz de recordar con todo detalle, y sin sufrir, aquel día en que Albertine se había marchado de tiendas con

El teórico PI podría afirmar ahora: «De nuestras tres actitudes ante el tiempo, una es irracional pero las otras dos están racionalmente requeridas. *Tenemos* que preocuparnos más por los dolores presentes, y *no podemos*, de una manera racional, preocuparnos por los dolores pasados, pero *no* tenemos que preocuparnos menos por los dolores que están en el futuro más lejano que por los que están en el futuro más próximo». Esta nueva concepción carece del atractivo de la generalidad. Había una simplicidad atractiva en la tesis de que meras diferencias en disposición temporal —meras respuestas a la pregunta «¿Cuándo?»— no pueden tener significación racional. Pero esta nueva concepción, aunque menos simple, puede estar justificada todavía. El teórico PI podría afirmar que, después de reflexionar, es intuitivamente plausible. Podría afirmar: «Cuando comparamos la condición de presente con la condición de pasado y con la distancia en el futuro, está claro que las dos primeras son por completo diferentes de la tercera. Las dos primeras tienen una significación racional obvia, justificando una diferencia en nuestro interés. Pero la tercera es obviamente trivial».

Estas intuiciones no son universales. De los que se sienten aliviados cuando los malos sucesos inevitables han sido pospuestos,

Françoise en vez de quedarse en el Trocadero. Lo rememoré con placer, como si perteneciera a un momento moral que no había conocido hasta entonces. Lo rememoré con todo detalle punto por punto, sin añadir ahora el más mínimo sufrimiento, más bien al contrario, de ese modo en que recordamos ciertos días de verano en los que tuvimos demasiado calor mientras transcurrían, y de los que sólo después que han pasado conseguimos extraer su puro patrón de fino oro y de azul celeste imperecedero».

O bien, Proust (1), p. 107:

«Una impresión de amor no guarda proporción con las demás impresiones de la vida, pero no es cuando se pierde entre éstas que podemos hacernos cargo de ella. No es desde su pie, en el tumulto de la calle y entre las casas apretadas, sino cuando estamos lejos, desde la falda de una colina vecina, a una distancia desde la que toda la ciudad ha desaparecido, o se muestra sólo como una confusa masa sobre la tierra, que podemos, en la tranquila distancia de la soledad y la oscuridad, apreciar la cima, única, persistente y pura, de una catedral.»

Sería posible vivir una vida de gran felicidad retrospectiva y anticipadora, aunque uno nunca encuentre, en el momento presente, ningún placer en ninguna de sus experiencias.

muchos no consideran que este alivio sea irracional. O consideremos otro efecto de la predisposición a favor de lo próximo: la creciente excitación que sentimos cuando se aproxima al presente un buen suceso —como ese momento en el teatro en el que se desvanecen las luces del patio de butacas—. Muchos juzgarían que esta excitación no es irracional.

El teórico PI podría decir: «Los que tienen estas intuiciones no han considerado suficientemente la cuestión. Los que *sí* la han considerado, como los filósofos, generalmente coinciden en que es irracional preocuparse más por el futuro más próximo».

Como he dicho, puede que la coincidencia de los filósofos no justifique su concepción. La teoría del Propio Interés ha sido dominante durante mucho tiempo. Como PI se ha enseñado durante más de dos milenios, tenemos que esperar encontrar algún eco en nuestras intuiciones. PI no puede justificarse simplemente con una apelación a intuiciones que su enseñanza puede haber producido.

Si el paso del tiempo no es una ilusión, el teórico PI no necesita apelar sólo a nuestras intuiciones. Puede afirmar que el paso del tiempo justifica la predisposición a favor del futuro. Si se le pide que explique por qué, tal vez lo encuentre difícil. No hay, por ejemplo, ninguna sugerencia de que el pasado sea irreal. Sería fácil ver por qué, si el pasado no fuese real, los dolores pasados no importarían. No es tan obvio por qué, puesto que el tiempo transcurre, los dolores pasados no importan.

El teórico PI podría afirmar: «Supongamos que aceptamos la metáfora de que el alcance de "ahora" se mueve en el futuro. Esto explica por qué, de las tres actitudes ante el tiempo, una es irracional y las otras dos son racionalmente requeridas. Los dolores importan sólo a causa de cómo se sienten cuando están en el presente, o bajo el alcance de "ahora". Por eso tenemos que preocuparnos más por nuestros dolores cuando los tenemos *ahora*. "Ahora" se mueve en el futuro. Por eso los dolores pasados no importan. Una vez que los dolores han pasado, únicamente se apartarán del alcance de "ahora". Las cosas son diferentes con la cercanía en el futuro. El paso del tiempo no justifica preocuparse más por el futuro pró-

ximo puesto que, por muy distantes que se hallen los dolores futuros, *se acabarán poniendo* dentro del alcance de "ahora"».

No está claro que estos sean buenos argumentos. El último, sobre todo, puede incurrir en una petición de principio. Pero el teórico PI podría en vez de esto afirmar que, al apelar al transcurso del tiempo, no necesitamos argumentos. Podría afirmar que, una vez más, no tenemos necesidad de explicación adicional. Puede ser otra verdad fundamental que, como el tiempo pasa, el sufrimiento pasado, simplemente, no puede importar —no puede ser el objeto de una preocupación racional—. Sintiempo no se sintió aliviado al saber que su suplicio estaba en el pasado. Puede que esto no implique el tipo de error que pueda explicarse. El error puede ser tan palmario que se halle más allá del alcance de la argumentación.

69. Una asimetría

Puede ocurrir que al abandonar la apelación a la neutralidad temporal y en su lugar apelar al paso del tiempo, el teórico del Propio Interés haya fortalecido su posición. Pero podríamos considerar una última clase de casos. Los llamo *El Sufrimiento Pasado o Futuro de Los que Quiero*.

> *Caso Uno.* Soy un exiliado de cierto país, donde he dejado a mi madre viuda. Aunque estoy muy preocupado por su situación, rara vez recibo noticias suyas. Me he enterado hace un tiempo de que está enferma terminal, y no le queda mucho de vida. Ahora me cuentan algo nuevo. La enfermedad de mi madre le produce grandes dolores, que las drogas no pueden aliviar. Los próximos meses, antes de que fallezca, tendrá que hacer frente a un suplicio terrible. Yo ya sabía que iba a morir pronto. Pero me quedo profundamente angustiado al enterarme del sufrimiento que va a tener que soportar.
>
> Un día después me cuentan que me habían informado mal en parte. Los hechos eran correctos, pero no su disposición temporal. Mi madre pasó muchos meses de sufrimiento, pero ahora ya está muerta.

¿Debería sentirme ahora notablemente aliviado? Había pensado que el suplicio de mi madre estaría en el futuro, pero estaba ya en el pasado. Según la nueva opinión del teórico PI, los dolores pasados simplemente no importan. El enterarme del sufrimiento de mi madre *no* me da ahora razón alguna para estar angustiado. Ahora es como si hubiera muerto sin dolor. Si todavía estoy angustiado, soy como Sintiempo. Estoy cometiendo un error tan palmario que está más allá del alcance de la argumentación.

Este último ejemplo puede sacudir al teórico PI. Puede que encuentre difícil de creer que mi reacción es irracional. Podría decir: «¿Cómo es posible que pueda importarte el que tu madre tuviera esos meses de sufrimiento? Aun si los tuvo, su sufrimiento *está en el pasado*. No son en absoluto malas noticias». Pero cuando las aplicamos a mi preocupación por otra persona, estas observaciones parecen menos convincentes.

En vez de esto, el teórico PI podría tratar de modificar su concepción. Podría decir: «No debería haber afirmado que los dolores pasados simplemente no importan. Lo que implica el paso del tiempo es que importan *menos*». Pero esta revisión no es justificable. Una vez que un dolor está en el pasado, ha pasado completamente. Estar en el pasado no es una cuestión de grado. No es plausible afirmar que, como el tiempo pasa, lo que es racional es tener *alguna* preocupación por el dolor pasado, pero *menos* que por el dolor futuro. ¿Y qué deberíamos decir de Mi Pasado Suplicio? En estos casos, considero mi sufrimiento pasado con absoluta indiferencia. ¿Es esto irracional? ¿Debería estar algo angustiado, aunque menos de lo que lo estoy por mi sufrimiento futuro? Una apelación al transcurso del tiempo no puede apoyar esta afirmación de manera convincente. Y es difícil de creer que, en estos casos, mi indiferencia sea irracional.

Mis ejemplos revelan una sorprendente asimetría en nuestra preocupación por nuestro propio pasado y por el de otras personas. No me sentiría en absoluto angustiado si se me recordara que yo mismo una vez tuve que soportar varios meses de sufrimiento. Pero me quedaría muy angustiado si me enterara de que, antes de morir, mi madre tuvo que soportar un suplicio semejante.

Esta asimetría se reduce en el

Caso Dos. Como el Caso Uno, salvo que, aunque mi madre sufrió durante varios meses, todavía está viva y ahora no tiene dolores.

Me sentiría menos angustiado aquí al enterarme del pasado sufrimiento de mi madre. Se puede explicar esta diferencia. Si mi madre es como yo, ahora contempla con indiferencia su suplicio pasado. (Podemos suponer que sus recuerdos de este suplicio no son en sí mismos penosos.) Si hay una asimetría en nuestra preocupación por nuestro propio sufrimiento pasado y por el sufrimiento pasado de otras personas, no sería sorprendente si fuese más clara en los casos en que los otros estuvieran ya muertos. Si mi madre está todavía viva, mi actitud presente será naturalmente afectada por lo que puedo suponer que es su actitud presente. Como resulta que puedo suponer que *ella* contempla ahora con indiferencia su sufrimiento pasado, esto podría reducir mi preocupación por su sufrimiento. Pero, si mi madre está ya muerta, no ve con indiferencia su sufrimiento pasado. Como mi preocupación por su sufrimiento pasado no puede ser afectada por su actitud presente, entonces es cuando mi preocupación se muestra en su forma más pura.

¿Supone alguna diferencia el que su sufrimiento durara hasta su muerte? Consideremos el

Caso Tres. Me entero de que mi madre sufrió durante varios meses, pero que, antes de que muriera, tuvo un mes libre de dolor. Hubo en su vida un período en que su sufrimiento había quedado en el pasado y por eso ya no le importaba a ella.

Si esto es lo que sé, ¿supondría una gran diferencia en mi preocupación? Creo que sería como mucho una pequeña diferencia. Me sentiría profundamente angustiado al enterarme de que mi madre sufrió durante esos meses, aunque también supiera que tuvo un mes en el que ese sufrimiento quedó en el pasado. Lo que me angustia no es sólo enterarme de la *muerte dolorosa* de mi madre. Si sólo fuera esto lo que me angustiara, y no me sintiera angustiado al enterarme de que ella tuvo que soportar un gran sufrimiento unos meses antes de morir, mi reacción sería tan especial que a lo mejor podría ignorarse. Pero mi preocupación por el pasado de las personas a las que quiero, y que ahora están muertas, no es meramente una preocupación por que no tuvieran muertes dolorosas. Me angustiaría al enterarme de que, en cualquier momento de sus vidas, tuvieron meses de dolor de los cuales yo no supe nada previamente. Creo que la mayoría de la gente, en este aspecto, es como yo.

Consideremos finalmente el

Caso Cuatro. Lo mismo que el Caso Tres, salvo que no me entero del sufrimiento de mi madre, puesto que supe de él en su momento.

Aunque siempre he tenido este conocimiento, yo seguiría entristeciéndome con el pensamiento de que, en la vida de mi madre, hubo varios meses de sufrimiento. Una vez más, pienso que una cosa similar se aplicaría a la mayor parte de la gente. Hay aún una sorprendente asimetría con nuestra actitud ante nuestro propio sufrimiento pasado, que la mayoría de nosotros contempla casi con indiferencia.

Puede objetarse: «Si trazamos distinciones, esta asimetría desaparece. Preguntas si, cuando ha quedado en el pasado, el sufrimiento *importa.* Esto mezcla cuestiones diferentes. Una cuestión es la de si debes *sentir simpatía*, y otra la de si debes *estar preocupado.* Si el sufrimiento queda en el pasado supone una diferencia, no en relación con la simpatía, sino sólo con la preocupación. Sentimos simpatía sólo por otras personas. *Por eso* tu ves tu pasado sufrimiento con indiferencia. No puedes simpatizar contigo mismo. Cuando te enteras del sufrimiento pasado de tu madre, sientes y debes sentir simpatía. Pero sería irracional *preocuparse* por este sufrimiento pasado, igual que sería irracional preocuparse por tu propio sufrimiento pasado» [43].

Estas tesis no eliminan, pienso yo, la asimetría. Al comienzo del Caso Uno, me dijeron que mi madre iba a sufrir varios meses antes de morir. Un día después el mensaje fue corregido: ella sufrió varios

[43] Esta objeción me fue sugerida por J. Broome, R. Swinburne y J. Thomson.

meses antes de morir. Según las tesis que se acaban de formular, yo debería estar muy preocupado el día anterior, cuando creía que el sufrimiento de mi madre iba a ser en el futuro. Cuando me entero de que quedaba ya en el pasado, debería dejar de estar preocupado, aunque todavía debería sentir simpatía. Cuando dejo de tener preocupación alguna, esto debería presumiblemente reducir mucho mi angustia, y además cambiar su cualidad. Pero estoy seguro de que, si ocurriera este caso imaginario, mi actitud no cambiaría en estas dos maneras. Yo podría estar algo menos angustiado, pero esta diferencia no sería grande. Ni tampoco cambiaría la cualidad de mi angustia.

El que un suceso quede en el pasado afectaría y debería afectar a aquellas de mis emociones que están vinculadas a actos posibles. Pero, en estos casos, cuando pienso que el sufrimiento de mi madre está en el futuro, no hay nada útil que pudiera hacer. Ni siquiera le puedo mandar un mensaje. No puedo por tanto tener el tipo de preocupación que es *activo*, el que busca modos en que puedo ayudar a la persona por la que estoy preocupado. En estos casos, mi preocupación sólo puede ser pasiva. Sólo puede ser tristeza y angustia, sin ningún impulso a buscar posibles remedios. Como mi angustia tomaría esta forma, su cualidad no cambiaría cuando me enterara de que el sufrimiento de mi madre queda en el pasado.

Admito que, cuando me enterase de este hecho, podría sentirme algo menos angustiado. Igual que mi preocupación podría ser afectada por la actitud de mi madre, si ella estuviera viva, podría ser afectada también por mi actitud hacia mi propio sufrimiento pasado. Este efecto puede eliminar la asimetría parcialmente. En mi preocupación por mi propio sufrimiento, supone toda la diferencia del mundo el que este sufrimiento quede en el futuro o en el pasado. No sería ninguna sorpresa si este hecho acerca de mis actitudes afectara a mi preocupación por el sufrimiento pasado de las personas que quiero. Como mi preocupación por su sufrimiento pasado no puede dejar de estar afectada por mi preocupación por mi propio sufrimiento pasado, mi preocupación por el sufrimiento de otros nunca puede tomar una forma completamente pura o no distorsionada. Y, como he afirmado, cuando me enterara de que el

sufrimiento de mi madre quedaba en el pasado, mi preocupación no disminuiría mucho.

Según la objeción presentada arriba, no me preocupa mi propio sufrimiento pasado porque uno no puede *simpatizar* consigo mismo. Esta tesis no hace nada para eliminar la asimetría. Es simplemente una redescripción. Concede que *hay* esta diferencia entre nuestras actitudes ante el sufrimiento pasado en nuestra propia vida y en la vida de aquellos a quienes queremos.

Esta asimetría hace más difícil defender la teoría del Propio Interés. Un teórico PI no puede defender convincentemente que esta asimetría esté racionalmente requerida. Sobre todo, no puede apelar aquí convincentemente al paso del tiempo. Si el paso del tiempo justifica mi absoluta indiferencia hacia mi propio sufrimiento pasado, o incluso hace de esta indiferencia un requisito racional, el teórico PI tiene que afirmar lo mismo sobre mi preocupación por aquellos que quiero. También es verdadero, en el caso imaginario de mi madre muerta, que su sufrimiento queda en el pasado.

¿Qué debería mantener el teórico PI acerca de nuestras actitudes ante el sufrimiento pasado? Podría afirmar: «No hay, aquí, una única actitud racional. Si contemplas tu propio sufrimiento pasado con total indiferencia, esto no es irracional. Pero tampoco lo sería si el conocimiento de tu propio sufrimiento pasado te causara gran angustia. De manera similar, no sería irracional el que te sintieras tremendamente angustiado al tener noticia del sufrimiento pasado de tu madre. Pero tampoco sería irracional si contemplaras su sufrimiento con total indiferencia».

Si el teórico PI admite como no irracional esta gama de actitudes diferentes hacia el pasado, ¿cómo puede defender su tesis de que, en nuestra preocupación por el futuro, debemos ser temporalmente neutrales? Todavía tiene que afirmar esta tesis. Pero si, en el caso del sufrimiento pasado, no sería irracional ni preocuparse en la misma medida, ni preocuparse menos, ni preocuparse en absoluto, ¿por qué en el caso del sufrimiento futuro hay sólo una actitud que es racional? Aunque no haya una inconsistencia absoluta, es difícil de creer una concepción que es tan permisiva en sus afirmaciones

acerca de una gama concreta de actitudes ante el tiempo, pero tan estricta en su afirmación acerca de otra gama distinta.

70. CONCLUSIONES

Concluyo que sólo hay dos concepciones que pueda esperar defender un teórico del Propio Interés:

(1) Si el paso del tiempo es una ilusión, la neutralidad temporal no puede ser irracional. El teórico PI podría volver a su tesis de que tenemos que ser temporalmente neutrales. Entonces tiene que afirmar que es irracional sentirse aliviado cuando el sufrimiento se ha pospuesto, y también cuando está en el pasado. Si critica la predisposición a favor de lo cercano, *tiene* que criticar también la predisposición a favor del futuro. Si el paso del tiempo es una ilusión, tiene que estar de acuerdo en (a) que no sería irracional *carecer* de la predisposición a favor del futuro. No puede además afirmar (b) que *no* es irracional *tener* esta predisposición, y a la vez (c) que *es* irracional tener la predisposición a favor de lo cercano. No hay argumento con el que pudiera apoyar estas tres afirmaciones. Si no condena la predisposición a favor del futuro, no puede condenar la predisposición a favor de lo cercano con la tesis de que es mala para nosotros. La predisposición a favor del futuro también es mala para nosotros. Y la racionalidad de una actitud no depende de si es mala para nosotros. Hay una diferencia entre estas dos actitudes ante el tiempo: aunque podemos actuar directamente a partir de la predisposición a favor de lo cercano, no podemos actuar directamente a partir de la predisposición a favor del futuro. Pero esto no puede apoyar la tesis de que sólo la primera predisposición es irracional. El teórico PI no puede establecer que la predisposición a favor del futuro no es irracional *porque* no podamos actuar a partir de ella. Si apela a la neutralidad temporal, tiene por eso que afirmar que es irracional sentirse aliviado cuando nuestro sufrimiento queda en el pasado. Vamos a encontrar esto difícil de creer.

(2) Si el paso del tiempo no es una ilusión, el teórico PI podría defender una concepción diferente. Podría sostener que, puesto que el tiempo pasa, el sufrimiento pasado no puede importar. Puede entonces sostener que es irracional para Sintiempo no sentirse aliviado cuando tiene noticia de que su sufrimiento ha acabado. Esta concepción la encontraremos plausible cuando nos pongamos a pensar en nuestro propio pasado. Pero el teórico PI tiene también que sostener que, cuando me angustio al enterarme de que mi madre sufrió antes de morir, esto es irracional. Vamos a encontrar esto difícil de creer.

El mismo teórico PI es posible que encuentre esta última afirmación difícil de creer. Si la abandona, tiene que abandonar también su apelación al paso del tiempo. Mientras que esta apelación podría dar apoyo a la tesis excesiva de que el sufrimiento pasado simplemente no importa, no puede apoyar la tesis de que estamos racionalmente requeridos a tener *alguna* preocupación por el sufrimiento pasado, pero una preocupación *menor*. Ni tampoco puede demostrar que sea irracional la diferencia en nuestras actitudes hacia el sufrimiento en el caso de nuestro propio pasado y en el del pasado de otras personas.

Aun si el paso del tiempo no es una ilusión, el teórico PI podría volver a su primera concepción: el requisito de neutralidad temporal. Puede entonces condenar la predisposición a favor de lo próximo con la tesis de que una mera diferencia en la disposición temporal no puede tener significación racional. Puede defender que, aunque sea racionalmente significativo *quién* siente el dolor, no puede ser significativo *cuándo* se siente el dolor.

Si vuelve a esta concepción, el teórico PI tiene que condenar la predisposición a favor del presente. Fue aquí donde la neutralidad temporal dio la impresión de ser menos plausible. ¿Cómo puede ser irracional tener más en cuenta mi horrible dolor cuando lo estoy pasando? El teórico PI podría decir: «En un sentido, esto no es irracional. El dolor horroroso es malo sólo a causa de cuánto lo tienes en cuenta en el momento de tenerlo. Pero, en otro sentido, no deberías estar predispuesto a favor del presente. Sería irracional dejar que tal predisposición influya en tus decisiones. Aunque tengas más en cuenta el dolor horroroso en el momento de tenerlo, no deberías por esta causa poner fin a tu dolor horroroso presente al

coste previsto de un dolor horroroso mayor después. A nivel de primer orden, tienes más en cuenta el dolor horroroso en el momento de sentirlo. Pero no deberías estar más preocupado por su estar en el presente más bien que en el futuro. A nivel de segundo orden, donde tomas decisiones que afectan a la duración y a la disposición temporal de tu sufrimiento, puedes y debes ser temporalmente neutral».

Si lo que exige es neutralidad temporal, el teórico PI tiene que condenar también la inclinación hacia el futuro. Podría decir: «Sería lógico pensar que esta inclinación fue producida por la Evolución. Esto explica por qué se aplica sólo, o con más fuerza, a nuestras propias vidas. Cuando consideramos las vidas de los demás, podemos elevarnos por encima de nuestra herencia evolutiva, y podemos ver la plausibilidad de la neutralidad temporal».

Cuando alguna creencia o actitud tiene una explicación evolutiva, esto, en sí mismo, tiene implicaciones neutrales. No puede por sí mismo demostrar que esta creencia o actitud esté o no esté justificada. Pero supongamos que tenemos otras razones para cuestionar una actitud. Sus defensores es posible que digan entonces: «El hecho de que todos tengamos esta actitud es una razón para pensar que está justificada. ¿Por qué la habrían compartido tantos, si no estuviera justificada?». Al contestar a *esta* pregunta, una explicación evolutiva puede poner en tela de juicio lo que explica. Socava la explicación rival, la de que tenemos esa creencia o esa actitud *porque* está justificada. El teórico PI podría por ello sostener que nuestra predisposición a favor del futuro, en nuestra propia vida, es un *mero producto* de la Evolución, y no está racionalmente justificada. Y esta tesis parece venir apoyada por la asimetría en nuestra preocupación por nuestra propia vida pasada y por la vida pasada de otras personas [44].

El teórico PI tendría que aplicar esta tesis a Mis Operaciones Pasadas o Futuras. En estos casos yo querría que fuese verdadero que sufrí durante varias horas ayer, en vez de que sufriré durante una hora hoy dentro de un rato. El teórico PI tiene que afirmar de

[44] Véase Singer (2).

nuevo que esta preferencia es irracional, y que, en general, es irracional sentirse aliviado cuando nuestro sufrimiento está en el pasado. Aun dando por buena su nueva tesis sobre la Evolución, vamos a encontrar esto difícil de creer.

He descrito las dos concepciones que puede defender de la forma más convincente el teórico del Propio Interés. Cada una de estas concepciones incluye al menos una tesis que es difícil de creer. Esto es una debilidad de la teoría del Propio Interés. Y es una debilidad adicional que haya una elección entre estas concepciones. Puede que sea irracional preocuparse menos del futuro más lejano. Pero no podemos estar seguros de ello mientras no sepamos por qué razón.

9

POR QUÉ DEBEMOS RECHAZAR PI

Un teórico del Propio Interés tiene que condenar la predisposición a favor de lo próximo. Una objeción es que esta predisposición es irracional. Como venimos de ver, esta objeción no es pan comido.

71. La apelación a remordimientos posteriores

El teórico PI podría apelar a una objeción diferente. Podría decirle a Próximus:

> *Ahora* no lamentas tu predisposición a favor de lo próximo. Pero la *lamentarás*. Cuando pagues el precio —cuando sufras el dolor que has pospuesto al coste de hacerlo peor— desearás haberte preocupado menos por tu futuro más próximo. Lamentarás tener esta predisposición. Es irracional hacer aquello de lo que sabes te vas a arrepentir.

Tal y como está formulada, esta objeción es imprecisa. Cuando Próximus paga el precio, tal vez lamente haber tenido en el pasado su predisposición a favor de lo próximo. Pero esto no demuestra

que tenga que arrepentirse de tener esta predisposición ahora. Una afirmación similar se aplica a los que se guían por el propio interés. Cuando un hombre que se guía por el propio interés paga el precio que le impusieron los actos guiados por el propio interés de los demás, lamenta el hecho de que esas otras personas estén guiadas por su propio interés. Lamenta la predisposición que tienen a favor de sí mismas. Pero esto no le lleva a lamentar esta predisposición en sí mismo. Las verdades sobre PI a nivel interpersonal se aplican a P a nivel intertemporal. Igual que un hombre guiado por su propio interés lamenta su predisposición, no en sí mismo sino sólo en otros, Próximus lamenta su predisposición, no en sí mismo ahora, sino sólo en sí mismo en otras ocasiones. Cuando asumí que Próximus no lamenta su predisposición, bastó con asumir que no lamenta su predisposición *presente*. Esta es la predisposición a partir de la que siempre actúa. La objeción presentada arriba no demuestra que Próximus tenga que lamentar *esta* predisposición.

Podemos a continuación darnos cuenta de que Próximus no siempre lamentaría su predisposición pasada. En el pasado *más próximo*, los que ahora son su presente y su futuro cercano estaban los dos próximos; de forma que estaba entonces predispuesto a favor de ambos. Como ahora y durante algún tiempo se está aprovechando de esta disposición pasada, ahora estará *contento* de haberla tenido.

El teórico PI podría decir: «Cuando pagues el precio, lamentarás tu predisposición pasada. Como esto es cierto, *debes* ahora lamentar incluso tu predisposición presente. En el futuro lamentarás tu predisposición presente; y ahora te preocupas por tu futuro. Preocuparse por el propio futuro conlleva desear evitar aquello de lo que uno pueda arrepentirse. Como te arrepentirás de tu predisposición presente, ahora debes desear no tenerla».

Próximus podría contestar: «En el futuro más distante es verdad que lamentaré mi predisposición presente. Y esto me da razones para desear no tener esta predisposición ahora. Pero en el futuro más cercano estaré contento de haber tenido mi predisposición presente, puesto que entonces me beneficiará. Esto me da razones para estar contento de tener esta predisposición ahora».

El teórico PI podría decir: «Tu alegría en el futuro próximo será contrarrestada por tus remordimientos en el futuro más distante. Esto será así porque pospones los dolores al coste previsible de empeorarlos. Como tu predisposición presente hará que más adelante tengas más remordimientos que alegría, tus razones para desear ahora no tener esta predisposición son, de las dos razones rivales, las más fuertes».

Próximus podría responder: «Mis remordimientos futuros es verdad que, imparcialmente considerados, contrarrestarán mi alegría futura. Pero yo no considero mi futuro imparcialmente. Me interesa más lo que está próximo. Como mi alegría futura está más próxima, compensa mis remordimientos futuros».

El teórico PI podría contestar: «Es irracional no ser imparcial». Próximus podría contestar: «Esta respuesta es o ineficaz o suicida. Si me derrota a mí, te derrota a ti. Los que están guiados por su propio interés no son imparciales. Igual que yo estoy predispuesto a favor de lo próximo, ellos están predispuestos a favor de sí mismos».

72. Por qué una derrota para Próximus no es victoria para PI

Al defender su predisposición a favor de lo cercano, Próximus la compara con la predisposición a favor de uno mismo. El teórico PI podría decir:

Yo me preocupo menos por lo que les pasa a los demás. Tú te preocupas menos por lo que te ocurrirá a ti más tarde. Estos momentos posteriores *serán*, tarde o temprano, un *ahora* para ti. Pero las demás personas jamás serán un *yo* para mí. Así que tu analogía falla [45].

Esta objeción tiene cierta fuerza. Supongamos que, por esta o por otras razones, rechazamos lo que sostiene Próximus. Supon-

[45] Esta objeción, que Nozick me sugirió, tiene menos fuerza cuando se aplica a la predisposición a favor del futuro.

gamos que, a pesar de las dificultades señaladas en el capítulo 8, concluyéramos que la predisposición a favor de lo próximo *es* menos racional que la predisposición a favor de uno mismo. Como he explicado, esto *no* demostraría que debamos aceptar PI. La predisposición a favor de lo próximo es el blanco favorito de los teóricos del Propio Interés. En mi intento de defender esta predisposición, estaba cuestionando a PI en su punto más poderoso. Si este intento tiene éxito, PI estará totalmente derrotada. Pero no se sigue de aquí que si el intento falla PI gane.

La mejor versión de la teoría del fin Presente es la versión Crítica. Como dije, CP puede defender que estamos racionalmente requeridos a preocuparnos por nuestro propio interés, y de un modo temporalmente neutral. Según esta versión de CP, Próximus es irracional, puesto que es irracional estar predispuesto a favor de lo próximo. Si pensamos que Próximus es irracional, esto no es razón para aceptar PI en vez de esta versión de CP.

73. La apelación a la inconsistencia

Supongamos que, como requiere esta versión de CP, me cuido de mi propio interés de un modo temporalmente neutral. Pero esta no es mi preocupación dominante. Como tengo otros deseos que a veces son más intensos, a veces obro de manera que sé que va contra mi propio interés. El teórico PI podría volver a su apelación a remordimientos posteriores. Podría decir: «Como obras en contra de tu propio interés, tus remordimientos futuros contrarrestarán tu alegría futura. A diferencia de Próximus, *tú* te interesas en la misma medida por tu futuro entero. Debes por tanto admitir que tus actos son irracionales. Es irracional hacer lo que sabes que en general lamentarás».

Esta objeción asume que, siempre que obro contra mi propio interés, después me arrepentiré de mi acto. Esta asunción no está justificada. Actúo así porque, aunque me preocupa mi propio interés, me preocupan aún más otras cosas. Como mi propio interés no es lo que más me preocupa, no deberíamos asumir que después me arrepentiré de mis actos.

Probablemente haya casos en que me arrepienta de un acto pasado. Pero de aquí no se sigue que mi acto fuera irracional. Supongamos que obré como lo hice porque aceptaba un juicio de valor que ahora rechazo. Puede que ahora lamente mi acto porque he cambiado de opinión. Pero mi acto todavía puede haber sido racional, puesto que yo obré a partir de un juicio de valor que en aquella época aceptaba.

Algo similar ocurre cuando cambian mis deseos sin que cambie mi opinión. Supongamos que, en el pasado, obré en contra de mis intereses porque quería ayudar a unas personas que estaban en peligro. Tomé prestada una gran suma de dinero para dárselo, sabiendo que, para devolver ese préstamo, habría de trabajar muchos años en una profesión que detesto. Ahora quiero ayudar a otras personas que están en peligro. No creo que esas personas tengan más derecho a ser ayudadas; pero, como en este momento soy muy consciente de que están en peligro, son ellas las personas por las que me preocupo más. A causa de mi acto anterior, ahora no puedo ayudar a estas personas. Como no he devuelto el préstamo, no puedo volver a pedir prestado para dárselo a ellas. Por eso me arrepiento de mi acto anterior. Este acto iba en contra de mis intereses, y ahora lo lamento. Pero, considerando lo que entonces más quería, no era irracional. La apelación a remordimientos posteriores tal vez demuestre que Próximus es irracional. Pero no supone ninguna objeción a la versión de CP que estoy discutiendo ahora.

Según esta versión de CP, yo debería tener un interés temporalmente neutral por mi propio futuro, pero no hace falta que este sea mi interés dominante. Y, como todas las versiones, esta versión de CP le da un peso especial a mis fines presentes. Puede por tanto ser cuestionada con una objeción que Nagel formula contra la más tosca teoría Instrumental. Nagel escribe que, según esta teoría:

> «Puede que ahora yo tenga razones para hacer justo lo que vaya a asegurar el fracaso de mis intentos *racionales* futuros; puede que yo tenga razones para hacer lo que sé que más tarde tendré razones para tratar de deshacer, y tendré por tanto que tener un cuidado especial en tender trampas y obstáculos insuperables en el camino

de mi yo futuro. Un sistema con consecuencias como esta no sólo falla a la hora de exigir la más elemental consistencia de la conducta a través del tiempo, sino que de hecho agudiza las posibilidades de conflicto fundamentando las intrigas de un individuo contra su yo futuro en el aparato de la racionalidad» [46].

La «inconsistencia» que Nagel describe no es una inconsistencia teórica. La teoría Instrumental no hace en momentos diferentes afirmaciones inconsistentes sobre lo que es racional que alguien haga. Ni tampoco necesita alguien que crea en esta teoría, o en la teoría Crítica del fin Presente, cuestionar su propia racionalidad en otros momentos. Es así porque, según todas las versiones de P, las razones son relativas no sólo al agente sino también al momento en que se actúa.

Es cierto que, según P, puede ser racional para mí hacer ahora lo que más tarde será racional para mí deshacer. Se puede acusar a P de ser intertemporalmente contraproducente, aun en los límites de una vida individual. Esto puede ser cierto incluso de la versión de CP que nos exige preocuparnos de nuestro propio interés de un modo temporalmente neutral. Supongamos que tengo esa preocupación, pero también tengo otros deseos más intensos, y estos deseos son diferentes en momentos diferentes. En un momento determinado yo puedo o bien (1) hacer lo que vaya a realizar de la mejor manera mis deseos presentes, o bien (2) hacer lo que vaya a realizar, o a permitirme realizar, de la mejor manera todos mis deseos a lo largo de mi vida. Según P, yo debería siempre hacer (1) antes que (2). Como vimos en la Sección 34, puede ser cierto que, si siempre sigo P, haciendo (1) antes que (2), tendré menos éxito a lo largo del tiempo a la hora de realizar mis deseos en cada momento.

Un teórico del Propio Interés podría sostener aquí que PI bate a P aun en los términos de P. En su versión de Realización de Deseos, PI me dice que haga siempre (2) antes que (1). Si siempre actúo así, tendré más éxito a lo largo del tiempo a la hora de realizar mis deseos en cada momento.

[46] Nagel (1), p. 40.

Expliqué cómo podemos contestar a esto. Hay una objeción similar contra PI. Según PI, puede ser racional para mí hacer lo que es racional para ti deshacer. Igual que P puede ser intertemporalmente contraproducente, PI puede ser interpersonal o colectivamente contraproducente. Una comunidad de personas guiadas por su propio interés lo haría mejor, aun en términos del propio interés, si todas ellas siguieran, no la teoría del Propio Interés, sino alguna versión de la moralidad. Pero ser colectivamente contraproducente no es lo mismo, en el caso de PI, que ser contraproducente de una manera irreparable. Cuando PI es colectivamente contraproducente, todavía sale individualmente bien parada. Y como PI es una teoría de la racionalidad individual, todavía funciona en sus propios términos.

¿Me hace falta escribir este párrafo? Ser intertemporalmente contraproducente no es lo mismo, en el caso de P, que ser contraproducente de una manera irreparable. Cuando P es intertemporalmente contraproducente, todavía sale bien parada en cada momento concreto. Es todavía cierto que, si sigo P en cada momento concreto, hago en cada momento concreto lo que realizará de la mejor manera mis deseos presentes. Aun en estos casos, desde el punto de vista del agente en el momento de actuar, P sale bien parada. Como este es el punto de vista al que apela P, todavía funciona en sus propios términos.

Estas dos objeciones no pueden refutar ni a P ni a PI. Pero las dos tienen cierta fuerza. A diferencia de P, PI no puede ser directa e intertemporalmente contraproducente. Esto le da a PI, en comparación con P, un cierto atractivo teórico. Y esto puede persuadirnos de que abandonemos P y aceptemos PI. Pero PI puede ser directa y colectivamente contraproducente. Esto no ocurre con una moralidad neutral respecto del agente. Lo cual le da a tal moralidad un atractivo teórico similar. Y esto puede persuadirnos de que demos el paso ulterior similar, de PI a una moralidad así.

Considerando la analogía entre las dos objeciones, la objeción contra P no da apoyo a PI. Estas objeciones entre las dos dan apoyo al Neutralismo. Si las objeciones salieran triunfantes, deberíamos

rechazar PI. Si las objeciones fallaran, no tendríamos razón para rechazar P [47].

Un teórico PI podría negar que la analogía tenga estas implicaciones. Podría afirmar que, aunque es verdad que una teoría aceptable no puede ser directa e intertemporalmente contraproducente, sí que puede ser directa e interpersonalmente contraproducente. ¿Cómo puede haber esta diferencia? El teórico PI tiene que sostener que la relación entre diferentes personas es, en los sentidos relevantes, diferente de la relación entre una persona individual en un momento y ella misma en otros momentos.

Estas relaciones son diferentes, en la mayoría de los aspectos. Mi relación contigo no es como la relación entre yo ahora y yo mismo mañana, o yo mismo dentro de cincuenta años. Pero estas

[47] Aunque tienen cierta fuerza, ninguna de las dos objeciones debería exagerarse. PI no impone una guerra de todos contra todos, por al menos tres razones. Los intereses, aun de las personas puramente egoístas, coinciden en una gran medida. Aunque estos intereses entren en conflicto, PI por sí misma puede decirles a esas personas que traten de resolver el conflicto, dando lugar a lo que llamo una solución política. (PI les dirá esto a esas personas cuando el caso involucre a un número suficientemente pequeño de personas.) Y la mayor parte de los que se guían por su propio interés no son enteramente egoístas, sino que tienen alguna preocupación por los demás.

Cada uno de los tres puntos tiene su análogo en el caso de P. Lo que alguien desearía más en momentos diferentes, si conociera los hechos y pensara con claridad, coincidiría en una gran medida. Aun cuando esto no fuera así, P mismo le dice a cada uno de nosotros que trate de resolver «peleas intertemporales» —que trate de conseguir consigo mismo en otros momentos soluciones políticas—. Si hay dos maneras en que alguien puede ahora tratar de realizar sus deseos presentes, y una de estas tiene más probabilidades de interferir con sus posteriores esfuerzos de realizar sus deseos futuros, P le dice a esta persona que actúe de la otra manera. Esto sería así incluso en el caso extremo en que alguien no tenga *ningún* interés en su propio futuro, y estemos apelando a la versión no crítica de P, la Teoría Instrumental. Como aprendí de G. Harman, aun en este caso, esta persona tendrá más éxito en conseguir lo que ahora desea si elige la manera de actuar que no va a interferir con sus esfuerzos futuros. Tendrá más éxito porque no tendrá después, en ese caso, razones que interfieran con los efectos de sus esfuerzos presentes. En tercer lugar, la mayoría de los agentes están de hecho interesados en sí mismos en momentos futuros. Y la teoría Crítica del fin Presente puede exigir tal interés, de una forma temporalmente neutral.

relaciones puede que, con todo, sean similares en los aspectos *relevantes*. Igual que un acto tiene que ser de un agente particular, tiene que hacerse en un momento concreto. Y muchas tesis sobre la racionalidad son verdaderas sólo cuando se aplican a una persona en un momento determinado. Dejan de ser verdaderas cuando se formulan para abarcar o la relación entre personas diferentes o la relación entre una persona en un momento determinado y la misma persona en otros momentos. Así que podría afirmarse, «Diferentes personas pueden tener racionalmente un conjunto de creencias inconsistentes, pero una persona individual no lo puede tener de manera racional». Pero esto no es correcto. Una persona individual puede tener racionalmente creencias inconsistentes si las tiene en momentos diferentes. La persona es irracional sólo si las tiene al mismo tiempo. Lo mismo ocurre con las preferencias intransitivas: preferir X a Y, Y a Z y Z a X. Como se ha señalado a menudo, tres personas pueden cada una de ellas tener una de estas preferencias sin ser irracionales. Pero también puede una persona individual, si las tiene en momentos diferentes. Como estas afirmaciones sugieren, cuando consideramos tanto la racionalidad teórica como la práctica, la relación entre una persona ahora y ella misma en otros momentos *es* similar de forma relevante a la relación entre diferentes personas.

74. Conclusiones

En los capítulos del 6 al 8 presenté varios argumentos contra la teoría del Propio Interés. Estos argumentos justifican una de dos conclusiones. Puede que los argumentos demuestren sólo que PI no puede batir a P. Según esta conclusión, la disputa entre estas teorías acaba en un empate, en tablas. Cuando PI y P entran en conflicto lo racional sería seguir a una de ellas. Pero, como explico en el Apéndice B, esta conclusión equivaldría en la práctica a una derrota para PI.

La otra conclusión es que debemos rechazar PI. Tanto en la práctica como en la teoría, esto sería, para PI, una completa derrota. (Si rechazamos PI, como explico en el Apéndice C, esto puede

afectar a nuestra elección entre las diferentes teorías sobre el propio interés.)

Creo que mis argumentos justifican esta conclusión más contundente. Comencé con una metáfora estratégica. La teoría del Propio Interés tiene dos rivales: la moralidad y la teoría del fin Presente. En algunos aspectos, ocupa un lugar intermedio entre estas dos rivales. Por eso es vulnerable de un modo que es con frecuencia fatal: puede ser atacada desde dos direcciones. La teoría del Propio Interés ha sido la teoría dominante durante mucho tiempo en nuestra tradición intelectual. Pero este dominio se ha derivado en gran medida del fracaso de sus dos rivales en atacar juntas. Cuando recibía el ataque de los teóricos morales, robaba fuerza de la teoría del fin Presente, y viceversa.

Cuestioné la teoría del Propio Interés desde ambas direcciones. Esto aseguró que PI fuese juzgada sólo según sus propios méritos. Evité el caso engañoso en que lo que una persona hace sólo le afecta a ella misma. PI le dice a esta persona que haga todo lo que sería mejor para ella. Como ella es la única persona a quien afectan sus actos, hace lo que es mejor para todos los afectados. Hace lo que, considerado imparcialmente, tiene los mejores efectos. En tal caso, PI coincide con la benevolencia imparcial. PI puede ser mejor juzgada cuando estos dos entran en conflicto: cuando lo que es mejor para el agente sería peor, y por un margen más amplio, para otras personas. Como estos casos demuestran, PI insiste en un patrón de intereses sesgado. PI no es sólo prudencia, sino también egoísmo. Insiste en que un agente racional tiene que dar importancia suprema a su propio interés, *cualquiera que sea* el coste para los demás.

Entonces cuestioné PI desde la otra dirección. Consideré casos en que PI entra en conflicto con la teoría del fin Presente. En estos casos, aunque el agente conoce los hechos y piensa con claridad, *no quiere* dar importancia suprema a su propio interés. PI sostiene que, cualquiera que sea el coste para los demás, un agente racional *tiene que* estar predispuesto a su propio favor, aunque, en un momento de serenidad, ni tenga ni quiera tener esta predisposición.

¿Es plausible esta tesis? ¿Es esta predisposición la única racional, la racional por excelencia? ¿Es todo deseo o interés diferente

menos racional? *Esta es la cuestión central.* Mi Primer Argumento contesta No. Defiendo que, comparados con la predisposición a favor de uno mismo, hay otros diversos deseos que no son menos racionales. Un ejemplo es el deseo de actuar en interés de otras personas. Puede ser racional realizar ese deseo, aun cuando uno sepa que su acto va en contra de su propio interés. Otros ejemplos son ciertas clases de deseos de logro. Un creador puede querer que sus creaciones sean las mejores posibles. Un científico, un filósofo, pueden querer hacer algún descubrimiento o avance intelectual fundamental. Defiendo que estos y otros deseos no son menos racionales que la predisposición a favor de uno mismo. Si uno de estos es el deseo más intenso de alguien, una vez consideradas todas las cosas, sería racional para él hacer que se realice, aunque sepa que su acto va en contra de su propio interés.

La Primera Respuesta del teórico PI contradice estas afirmaciones. Esta respuesta sostiene que la predisposición a favor de uno mismo es la predisposición racional por excelencia. Como no puedo probar que tengo razón al rechazar esta tesis, mi Primer Argumento no es decisivo. Pero creo que sale triunfante. Creo que el teórico PI no tiene una buena respuesta que dar a este argumento. La predisposición a favor de uno mismo *no* es la predisposición racional por excelencia. Hay al menos un deseo que no es menos racional: el deseo de beneficiar a los demás. Como hay al menos un deseo de estos, debemos rechazar PI y aceptar alguna versión de CP.

La Segunda Respuesta del teórico PI apela a la tesis de que la fuerza de cualquier razón se extiende a través del tiempo. Según esta tesis, como *tendré* razones para tratar de realizar mis deseos futuros, tengo esas razones *ahora*. Las razones para actuar no pueden ser relativas a un momento particular. Un argumento a favor de esta tesis puede también demostrar que las razones para actuar no pueden ser relativas al agente. El argumento puede demostrar que la fuerza de cualquier razón se extiende a través de las vidas de diferentes personas. Esto es lo que se demostraría si el argumento de Nagel saliese triunfante. Esta conclusión haría fracasar tanto a PI como a P. Para evitar esta conclusión, el teórico PI tiene que defender que las razones pueden ser relativas al agente. Yo sostengo que,

si las razones pueden ser relativas, pueden ser relativas al agente en el momento de actuar. Como mostré en las Secciones de la 59 a la 61, puede ser cierto que yo tuve o tendré determinadas razones para actuar, aunque no las tenga ahora. Esto socava la Segunda Respuesta del teórico PI. Y mi apelación a la relatividad plena dio razones adicionales para rechazar PI.

El teórico PI tiene que defender también que es irracional preocuparse menos por el futuro distante de uno. El capítulo 8 demostró que, al defender esto, el teórico PI tiene que aceptar una de dos concepciones, cada una de las cuales con implicaciones que son difíciles de creer. Esto es otra objeción contra PI. Para los propósitos del argumento, asumí que esta objeción puede neutralizarse. Asumí que es irracional preocuparse menos por el futuro lejano. Esto no demuestra que debamos aceptar PI. Podríamos aceptar la versión Crítica de la teoría del fin Presente. Y CP puede defender que estamos racionalmente requeridos para preocuparnos por nuestro propio interés, de un modo temporalmente neutral. No es esta tesis lo que distingue a estas dos teorías.

PI exige que aceptemos una tesis mucho más contundente. No basta con que tengamos esta predisposición temporalmente neutral a nuestro favor. Tenemos que ser gobernados siempre por esta predisposición, cualquiera que sea el coste para los demás, y aunque ni la tengamos ni queramos tenerla. Esta tesis nos devuelve a la cuestión central. Según mi Primer Argumento, esta tesis exige la asunción de que esta predisposición es la predisposición racional por excelencia. Exige la asunción de que es irracional preocuparse más por cualquier otra cosa, como por ejemplo la moralidad o los intereses de otras personas. Debemos rechazar esta asunción. Si el teórico PI no tiene otra respuesta, debemos rechazar PI.

El teórico PI tiene otros dos argumentos: la apelación a remordimientos posteriores, y la apelación a la inconsistencia. Aunque estos argumentos tienen un atractivo intuitivo, no proporcionan respuestas para mi Primer Argumento. *Concluyo que debemos rechazar PI.* Como predije, la teoría del Propio Interés no puede sobrevivir a un ataque combinado de sus dos rivales: la teoría del fin Presente y la moralidad.

La mejor versión de la teoría del fin Presente es la versión Crítica. Recordemos además que, si aceptamos CP, *podríamos* defender que está racionalmente requerido el que nuestro deseo más intenso sea evitar obrar mal. He dejado abierto si *debemos* añadir esta tesis a CP. Como consecuencia, no hay *dos* teorías supervivientes acerca de la racionalidad. *Los teóricos morales deben aceptar CP.* No tienen razones para rechazar CP, puesto que CP puede dar a las razones morales toda la importancia que ellos piensan que deben tener.

Recordemos finalmente que *todas* las teorías posibles de la racionalidad son versiones de CP. Como esto es cierto, debemos aceptar CP, sea lo que sea lo que creamos. Este rasgo de CP puede parecer una debilidad que la convierte en una teoría vacía. Pero es un rasgo de fuerza. Podemos ver más claramente lo que asumen diferentes teorías cuando se reformulan como versiones de CP. Y, mientras que dejé abierto lo que CP debe defender sobre las razones morales, no dejé abiertas otras dos cuestiones. Consideremos a los seguidores de Hume, que niegan que los deseos puedan ser intrínsicamente irracionales o venir racionalmente requeridos. Si añadimos esta tesis a CP, llega a coincidir con PIn, la teoría puramente Instrumental. He defendido que debemos rechazar esta versión de CP. Hay patrones de intereses que son irracionales, y no nos dan ninguna razón para actuar. Y mi tesis principal es que debemos rechazar la versión de CP que coincide con PI. Debemos rechazar la asunción de que, comparado con la predisposición a favor de uno mismo, todo otro deseo es menos racional. Supongamos que nuestros deseos y juicios de valor, tanto individualmente como en conjunto, no son irracionales. Y supongamos que sabemos que lo que mejor realizará estos deseos va en contra de nuestro propio interés a largo plazo. Si esto es así, es *irracional* seguir PI. Es irracional hacer lo que va a favor de nuestro propio interés cuando sabemos que esto frustrará lo que, conociendo los hechos y pensando con claridad, queremos o valoramos más.

La mayoría de la gente ha creído en la teoría del Propio Interés durante más de dos milenios. Como consecuencia, puede parecer

absurdamente temerario defender que debemos rechazarla. ¿Cómo pueden cuatro capítulos derrocar el veredicto de la historia registrada? ¿Cómo puede haberse equivocado tanta gente? Hay dos respuestas.

(1) La mayoría asumió que, como tendremos una vida de ultratumba, o si no nos reencarnaremos, la moralidad y el propio interés coinciden siempre. Como tenían esta falsa creencia, estas personas pasaron por alto una de las objeciones a PI.

(2) Como con frecuencia ocurre cuando debemos rechazar una teoría, los que creyeron en ella no estuvieron completamente equivocados. Hay partes de PI que son plausibles. No es irracional preocuparse más por uno mismo. Y, en nuestra preocupación por nuestro propio interés, a lo mejor debemos ser temporalmente neutrales. La plausibilidad de estas tesis nos ayuda a explicar por qué tantas personas han creído en la teoría del Propio Interés. Pero estas tesis son también parte de una teoría más amplia, CP, que todos debemos aceptar. Consideremos esta analogía (demasiado grandiosa): Las Leyes de Newton son parcialmente correctas, pero ahora aceptamos una teoría diferente.

Tercera parte

LA IDENTIDAD PERSONAL

10

LO QUE CREEMOS SER

Entro en el teletransportador. Ya he estado antes en Marte, pero nada más que por el viejo método, un viaje en nave espacial que dura varias semanas. Esta máquina me enviará a la velocidad de la luz. Sólo tengo que apretar el botón verde. Como otros en mi situación, estoy nervioso. ¿Funcionará? Repaso lo que me han dicho que va a pasar. Cuando apriete el botón, perderé la conciencia y luego despertaré con la impresión de que sólo ha transcurrido un momento. En realidad habré estado inconsciente durante casi una hora. El escáner aquí en la Tierra destruirá mi cerebro y mi cuerpo, mientras registra los estados exactos de todas mis células. Entonces transmitirá esta información por radio. Viajando a la velocidad de la luz, el mensaje tardará tres minutos en llegar al replicador en Marte. Éste creará entonces, partiendo de materia nueva, un cerebro y un cuerpo exactamente como los míos. Será en ese cuerpo donde me despertaré.

Aunque creo que esto es lo que va a ocurrir, todavía vacilo. Pero entonces recuerdo cómo se reía mi mujer cuando, hoy al desayuno, le manifesté mi nerviosismo. Como me recordó, ella ha sido teletransportada a menudo, y nada va mal con *ella*. Aprieto el botón. Como se me pronosticó, pierdo la conciencia y aparentemente la recobró enseguida, pero en un cubículo diferente. Examinando mi nuevo cuerpo, no

encuentro ningún cambio en absoluto. Hasta está todavía en su sitio el corte que me hice en el labio superior esta mañana al afeitarme.

Pasan varios años durante los que soy teletransportado con frecuencia. Estoy otra vez en el cubículo, listo para otro viaje a Marte. Pero esta vez, cuando aprieto el botón verde, no pierdo la conciencia. Se escucha un zumbido, y luego el silencio. Salgo del cubículo y le digo al asistente: «No funciona. ¿Qué hice mal?».

«Sí que funciona», contesta, y me da una tarjeta impresa. Leo: «El nuevo escáner graba un cianotipo* de usted mismo sin destruir su cerebro ni su cuerpo. Esperamos que sepa apreciar las oportunidades que este avance técnico ofrece».

El asistente me cuenta que soy una de las primeras personas que usan el nuevo escáner. Añade que si me quedo una hora podré usar el intercomunicador para verme y hablar conmigo en Marte.

«Un momento», contesto, «Si estoy aquí no puedo estar *también* en Marte.

Alguien tose con mucha cortesía, un hombre de bata blanca que me pide hablar en privado conmigo. Nos vamos a su despacho, me dice que me siente, y hace una pausa. Luego dice: «Me temo que tenemos problemas con el nuevo escáner. Graba su cianotipo con la misma perfección y exactitud, ya lo podrá comprobar cuando se vea y hable consigo mismo en Marte. Pero parece que resulta nocivo para el sistema cardiaco cuando lo explora. A juzgar por los resultados que hemos tenido hasta ahora, aunque estará usted en Marte con una salud perfecta, aquí en la Tierra tiene que esperar un ataque cardíaco en los próximos días».

Después me llama el asistente por el intercomunicador. En la pantalla me veo a mí mismo justo igual que en el espejo por las mañanas. Pero hay dos diferencias. En la pantalla no aparece mi imagen invertida de derecha a izquierda. Y mientras que aquí estoy sin decir palabra, puedo ver y oír cómo empiezo a hablar, en el estudio de Marte.

* *Blueprint:* impresión fotográfica en blanco sobre fondo azul utilizada sobre todo para copiar dibujos mecánicos y planos arquitectónicos. «Cianotipo» se ha generalizado para referirse a «croquis» o «estructura». También está próximo a la idea de programa informático, en este contexto de un experimento de pensamiento que nos sitúa en el ámbito del funcionalismo computacional como teoría de la mente. (N. del t.)

¿Qué podemos aprender de esta historia imaginaria? Hay quienes piensan que podemos aprender poco. Esta habría sido la opinión de Wittgenstein [1]. Y Quine escribe: «El método de la ciencia ficción tiene sus usos en filosofía, pero... Me pregunto si son tenidos en cuenta apropiadamente los límites de este método. Buscar lo que "se requiere lógicamente" para la mismidad de la persona bajo circunstancias sin precedentes es sugerir que las palabras tienen una fuerza lógica que va más allá de aquella con que nuestras pasadas necesidades las han investido [2]».

Esta crítica podría estar justificada si no tuviéramos reacciones cuando consideramos esos casos imaginarios. Pero esos casos despiertan en la mayoría de nosotros poderosas creencias. Y no se trata de creencias sobre nuestras palabras, sino sobre nosotros mismos. Al considerar esos casos, descubrimos lo que pensamos que está implicado en nuestra propia existencia continua, o qué es lo que nos hace a nosotros ahora y a nosotros el año que viene la misma persona. Descubrimos nuestras creencias acerca de la naturaleza de la identidad personal a través del tiempo. Aunque nuestras creencias se revelan con la mayor claridad cuando consideramos casos imaginarios, también cubren casos reales, y nuestras propias vidas. En la Tercera Parte de este libro sostendré que algunas de estas creencias son falsas, pasando después a sugerir cómo y por qué importa esto.

75. El teletransporte simple y el caso de la línea secundaria

Al principio de mi historia, el escáner destruye mi cerebro y mi cuerpo. Mi cianotipo se transmite a Marte, donde otra máquina hace una *Réplica* orgánica mía. Mi Réplica piensa que es yo, y parece recordar haber vivido mi vida hasta el momento en que apreté el botón verde. En todos los demás aspectos, tanto físicos como psi-

[1] Véase, por ejemplo, *Zettel*, ed. Por G. Anscombe y G. Von Wright, y traducido por G. Anscombe, Blackwell, 1967, Proposición 350: «Es como si nuestros conceptos conllevaran un andamiaje de hechos... Si imaginas ciertos hechos de otra manera... entonces ya no puedes imaginar la aplicación de ciertos conceptos».

[2] Quine (1), p. 490.

colágicos, somos exactamente iguales. Si regresara a la Tierra, todos creerían que era yo.

El teletransporte simple recién descrito aparece a menudo en la ciencia ficción. Y algunos lectores de este tipo de literatura simplemente piensan que es la forma más rápida de viajar. Piensan que mi Réplica *sería yo*. Otros lectores de ciencia ficción, y algunos de los personajes de esta historia, adoptan una opinión diferente. Piensan que cuando aprieto el botón verde muero. Mi Réplica es *otra persona*, que ha sido fabricada para ser exactamente como yo. Esta segunda opinión parece venir apoyada por el final de mi historia. El nuevo escáner no destruye ni mi cerebro ni mi cuerpo. Aparte de recoger la información, se limita a dañar mi corazón. Mientras estoy en el cubículo, con el botón verde apretado, no parece ocurrir nada. Salgo y me entero de que en pocos días moriré. Luego hablo con mi Réplica en Marte, por televisión de doble sentido. Sigamos con la historia. Como mi Réplica sabe que estoy a punto de morir, trata de consolarme como los mismos pensamientos con los que hace poco intenté consolar a un amigo moribundo. Es triste darse cuenta, cuando a uno le llega el fin, de lo poco que consuelan estos pensamientos. Mi Réplica entonces me asegura que seguirá con mi vida donde yo la dejé. Ama a mi mujer, y entre los dos cuidarán de mis hijos. Y terminará el libro que estoy escribiendo. Además de tener todos mis borradores, tiene todas mis intenciones. Tengo que admitir que puede terminar mi libro tan bien como podría yo. Todas estas cosas me consuelan un poco. Morir cuando sé que tendré una Réplica no es tan malo como morir, simplemente. Aun así, pronto perderé la conciencia, para siempre.

En el teletransporte simple, me destruyen antes de ser replicado. Esto hace más fácil creer que *es* un modo de viajar —que mi Réplica *es* yo—. Al final de mi historia, mi vida y la de mi Réplica se solapan. Llamemos a esto el *caso de la línea secundaria*. En este caso, no puedo esperar viajar en la *línea principal*, despertando en Marte con cuarenta años de vida ante mí. Me quedaré en la línea secundaria, aquí en la Tierra, una línea que termina unos pocos días después. Como puedo hablar con mi Réplica, parece claro que *no* es yo. Aunque es exactamente como yo, ella es una persona y yo soy otra. Cuando me pellizco, no siente nada. Cuando me dé el ataque al corazón, tampoco sentirá nada. Y cuando yo muera, vivirá otros cuarenta años.

Si pensamos que mi Réplica no es yo, es natural asumir que mi futuro, en la línea secundaria, es casi tan malo como la muerte corriente. Voy a negar esta asunción. Como defenderé después, ser destruido y replicado es casi tan bueno como la supervivencia corriente. Pero podré defender esta tesis mucho mejor, junto con la concepción más amplia de la que es parte, después de discutir el pasado debate de la identidad personal.

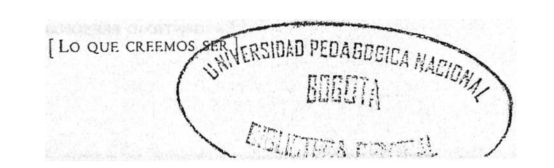
76. Identidad cualitativa e identidad numérica

Hay dos clases de igualdad o identidad. Yo y mi Réplica somos *cualitativamente idénticos*, o exactamente iguales. Pero puede que no seamos *numéricamente idénticos*, o una y la misma persona. De forma similar, dos bolas de billar blancas no son idénticas numéricamente pero pueden ser cualitativamente idénticas. Si yo pinto de rojo una de estas bolas, dejará de ser lo que era, cualitativamente idéntica consigo misma. Pero la bola roja que veo a continuación y la bola blanca que pinté de rojo son numéricamente idénticas. Son una y la misma bola.

Se podría decir de alguien, «Después de su accidente ya no es la misma persona». Esta es una afirmación acerca de los dos tipos de identidad. Decimos que *él*, la misma persona, *no* es ahora la misma persona. Esto no es una contradicción. Simplemente queremos decir que el carácter de esta persona ha cambiado. Esta persona numéricamente idéntica es ahora cualitativamente diferente.

Cuando nos preocupa nuestro futuro, es nuestra identidad numérica lo que nos preocupa. Puede que piense que después de mi boda no seré la misma persona. Pero esto no convierte a la boda en la muerte. Por mucho que cambie, todavía viviré si hay alguna persona viva que *será* yo.

Aunque nuestra preocupación principal es nuestra identidad numérica, los cambios psicológicos importan. En verdad, según cierta concepción, determinadas clases de cambio cualitativo destruyen la identidad numérica. Si me ocurren ciertas cosas, podría ser cierto no que me convirtiese en una persona muy diferente:



podría ser cierto que yo dejara de existir —que la persona resultante fuese alguien otro.

77. El criterio físico de identidad personal

Ha habido mucho debate sobre la naturaleza tanto de las personas como de la identidad personal a través del tiempo. Servirá de ayuda distinguir estas preguntas:

(1) ¿Cuál es la naturaleza de una persona?
(2) ¿Qué es lo que hace que una persona en dos momentos temporales diferentes sea una y la misma persona? ¿Qué está necesariamente implicado en la existencia continua de cada persona a través del tiempo?

La respuesta a (2) puede tomar esta forma: «*X* hoy es una y la misma persona que *Y* en un momento pasado *si y sólo si*...». Tal respuesta establece las *condiciones necesarias y suficientes* para la identidad personal a través del tiempo.

Al responder a (2) también contestaremos en parte a (1). Los rasgos necesarios de nuestra existencia continua dependen de nuestra naturaleza. Y la respuesta más simple a (1) es que, para ser una persona, un ser tiene que ser autoconsciente, consciente de su identidad y de su existencia continua a través del tiempo.

También podemos preguntar

(3) ¿Qué conlleva efectivamente la existencia continua de cada persona a través del tiempo?

Como nuestra existencia continua tiene rasgos que no son necesarios, la respuesta a (2) es sólo parte de la respuesta a (3). Por ejemplo, tener el mismo corazón y el mismo carácter no son necesarios para nuestra existencia continua, pero usualmente son parte de lo que esta existencia conlleva.

Muchos escritores utilizan la ambigua expresión «el criterio de identidad a través del tiempo». Algunos se refieren con esto a «nuestro modo de decir si algún objeto presente es idéntico a un objeto pasado». Pero yo me referiré a *lo que esta identidad necesariamente conlleva o a aquello en lo que consiste.*

En el caso de la mayor parte de los objetos físicos, según lo que llamo la *concepción estándar*, el criterio de identidad a través del tiempo es la continuidad física espacio-temporal de ese objeto. Esto es algo que todos comprendemos, aunque fracasemos en comprender la descripción que ahora daré. En el caso más simple de continuidad física, como el de las Pirámides, un objeto aparentemente estático continúa existiendo. En otro caso simple, como el de la Luna, un objeto se mueve de una manera regular. Muchos objetos se mueven de modos menos regulares, pero aun así trazan sendas espacio-temporales físicamente continuas.

Supongamos que la bola de billar que pinté de rojo es la misma que la bola blanca con la que el año pasado hice una jugada ganadora. Según la opinión estándar, esto es verdadero sólo si esta bola trazó tal senda continua. Tiene que ser cierto (1) que hay una línea a través del espacio y el tiempo que comienza donde estaba la bola blanca antes de que yo hiciera mi jugada ganadora y que termina donde está ahora la bola roja, (2) que en cada uno de los puntos de esta línea hubo una bola de billar, y (3) que la existencia de una bola en cada uno de los puntos de la línea fue en parte causada por la existencia de la bola en el punto inmediatamente precedente [3].

Hay cosas de ciertas clases que continúan existiendo aunque su continuidad física conlleve grandes cambios. Una «Camberwell Beauty» es primero un huevo, después una oruga, luego una crisálida y después una mariposa. Son cuatro fases en la existencia físicamente continua de un mismo organismo. Y hay otras clases de cosas que no pueden sobrevivir a cambios tan grandes. Supongamos que un artista pinta un auto-retrato y luego, pintando encima, lo convierte en el retrato de su padre. Aunque los dos retratos son más parecidos que una oruga y una mariposa, no son fases en la exis-

[3] Esto establece una condición necesaria para la existencia continua de un objeto físico. Saul Kripke ha sostenido, en conferencias, que esta condición no es suficiente. Como no asistí a esas conferencias, no puedo discutir este argumento.

tencia continua de una misma pintura. El auto-retrato es una pintura que el artista destruyó. En una discusión general sobre la identidad, tendríamos necesidad de explicar por qué el requisito de la continuidad física difiere de tales modos para diferentes clases de cosas. Pero aquí podemos ignorar este extremo.

¿Puede haber vacíos en la existencia continua de un objeto físico? Supongamos que tengo el mismo reloj de oro que le dieron a un muchacho, aunque quede desmontado durante un mes sobre la estantería de un relojero. Según una opinión, en la senda espacio-temporal trazada por este reloj no hubo un reloj en cada uno de los puntos, de modo que mi reloj no tiene una historia de continuidad física plena. Pero durante el mes que mi reloj estuvo desmontado, y no existió, todas sus partes tuvieron historias de plena continuidad. Según otra opinión, aun cuando estuvo desmontado, mi reloj existía.

Otra complicación afecta de nuevo a la relación entre una cosa compleja y las diversas partes de que está compuesta. Ocurre con algunas de estas cosas, aunque no con todas, que su existencia continua no necesita llevar consigo la existencia continua de sus componentes. Supongamos que reparamos de vez en cuando un barco de madera mientras está amarrado en el puerto, y que después de cincuenta años no contiene ninguno de los trozos de madera de los que al principio estuvo construido. Es todavía el mismo barco, porque, como barco, ha manifestado durante esos cincuenta años una continuidad física total. Así es, a pesar del hecho de que ahora está compuesto de pedazos de madera muy diferentes. Tales pedazos podrían ser cualitativamente idénticos a los originales, pero no son los mismos. Algo parecido ocurre en parte con el cuerpo humano. A excepción de algunas células cerebrales, las de nuestro cuerpo son reemplazadas por células nuevas varias veces en la vida.

Ahora acabo de describir la continuidad física que, según la opinión estándar, hace de un objeto físico el mismo objeto después de muchos días o años. Esto me va a permitir formular una de las concepciones que rivalizan en el debate de la identidad personal. Según esta concepción, lo que hace de mí la misma persona a través del tiempo es que tengo el mismo cerebro y el mismo cuerpo. El criterio de mi identidad a través del tiempo —o de lo que implica tal

identidad— es la continuidad física, a través del tiempo, de mi cerebro y mi cuerpo. Yo voy a seguir existiendo si y sólo si este cerebro y este cuerpo concretos siguen los dos existiendo como el cerebro y el cuerpo de una persona viva.

Esta es la versión más simple de esta concepción. Hay una versión mejor. Es esta

> *El criterio físico*: (1) Lo que resulta necesario no es la existencia continua del cuerpo entero, sino la existencia continua de *bastante* cerebro como para ser el cerebro de una persona viva. *X* hoy es una y la misma persona que *Y* en un momento pasado si y sólo si (2) suficiente cerebro de Y sigue existiendo, y es ahora el cerebro de X, y (3) esta continuidad física no ha tomado una forma «ramificada». (4) La identidad personal a través del tiempo consiste justamente en que se den hechos como (2) y (3).

(1) es claramente verdadero en ciertos casos reales. Hay personas que continúan existiendo aunque pierdan gran parte de su cuerpo, o el uso de su cuerpo. Explicaremos (3) más adelante.

Los que creen en el criterio físico rechazarían el teletransporte. Pensarían que no es una forma de viajar, sino una forma de morir. También rechazarían como inconcebible la reencarnación. Creen que nadie puede vivir después de la muerte como no viva esa vida en la resurrección del mismísimo cuerpo, físicamente continuo. Por eso hay cristianos que insisten en ser enterrados. Creen que si, al modo de los héroes griegos y troyanos, fueran incinerados sobre piras funerarias, y sus cenizas esparcidas al viento, ni siquiera Dios sería capaz de traerlos de nuevo a la vida. Dios podría crear tan sólo una Réplica, alguien diferente que fuera exactamente como ellos. Otros cristianos creen que Dios podría resucitar*los* si Él recompusiera sus cuerpos a partir de los pedazos de materia que, la última vez que estuvieron vivos, componían sus cuerpos. Lo cual sería como volver a montar mi reloj de oro [4].

[4] Según esta concepción, podría ser fatal vivir en lo que ha sido por mucho tiempo un área densamente poblada, como Londres. Aquí puede ser verdadero de muchos pedazos de materia que fueron parte de los cuerpos de muchas personas

78. El criterio psicológico

Hay quienes creen en un tipo de continuidad psicológica que se parece a la continuidad física. Es el que implica la existencia continua de una *entidad* o cosa puramente mental —un alma o sustancia espiritual—. Volveré después a esta concepción. Pero antes explicaré otro tipo de continuidad psicológica. Se parece menos a la continuidad física porque no consiste en la existencia continua de una entidad, pero implica sólo hechos que nos son familiares.

La que se ha discutido más es la continuidad de memoria, porque es la memoria la que nos hace a la mayoría de nosotros conscientes de nuestra propia existencia continua a través del tiempo. La excepción la constituyen las personas que sufren de amnesia. La mayor parte de los amnésicos pierden sólo dos conjuntos de recuerdos. Pierden todos sus recuerdos de experiencias pasadas concretas —o, para abreviar, sus *recuerdos experienciales*—. Y también pierden algunos de sus recuerdos de hechos, los que tratan de su propia vida pasada. Pero recuerdan otros hechos, y recuerdan cómo hacer diversas cosas, como hablar o nadar, por ejemplo.

Locke sugirió que la memoria experiencial proporciona el criterio de identidad personal [5]. Aunque esta no sea, como tal, una tesis convincente, pienso que puede ser parte de una tesis convincente. Por eso trataré de responder a algunos de los críticos de Locke.

Locke afirmaba que nadie puede haber cometido un crimen a no ser que recuerde ahora haberlo hecho. Podemos comprender que haya reticencias a castigar a la gente por crímenes que no puede recordar. Pero, tomada como una concepción sobre lo que está involucrado en la existencia continua de una persona, la afirmación de Locke es claramente falsa. Si fuera verdadera, sería imposible que

diferentes, cuando aún estaban vivas. No podrían ser resucitadas todas estas personas, puesto que no habría bastante de esta materia para ser juntada de nuevo. Hay quienes mantienen una versión de esta concepción que evita este problema. Son de la opinión de que un cuerpo resucitado necesita contener sólo una partícula del cuerpo original.

[5] Locke, Capítulo 27, Sección 16.

nadie olvidara ninguna de las cosas que hizo alguna vez, o ninguna de las experiencias que alguna vez tuvo. Pero esto *sí es* posible. Ahora no puedo acordarme de que me puse la camisa esta mañana.

Hay varios modos de ampliar el criterio de la memoria experiencial para que incluya estos casos. Recurriré al concepto de una cadena parcialmente superpuesta de recuerdos experienciales. Digamos que, entre X hoy e Y hace veinte años, hay *conexiones directas de memoria* si X puede acordarse ahora de haber tenido algunas de las experiencias que tuvo Y hace veinte años. Según la concepción de Locke, sólo esto es lo que hace a X e Y la misma persona. Pero aunque *no* haya esas conexiones directas de memoria, puede haber *continuidad de memoria* entre X ahora e Y hace veinte años, si entre X ahora e Y en aquel entonces ha habido una cadena parcialmente superpuesta de recuerdos directos. En el caso de la mayoría de los adultos, habría una cadena así. Cada día de los últimos veinte años la mayoría de nosotros recordaba algunas de sus experiencias del día anterior. Según esta versión revisada de la tesis de Locke, una persona actual X es la misma que una persona pasada Y si hay entre ellas continuidad de memoria.

Esta revisión anula una objeción a la tesis de Locke. También debemos revisar su tesis para que haga referencia a otros hechos. Además de recuerdos directos, hay diversos tipos distintos de conexión psicológica directa, como por ejemplo la que se da entre una intención y el acto subsiguiente en que la intención se realiza. Y otras conexiones directas de este tipo son las que se dan cuando una creencia, un deseo, o bien otro estado psicológico, siguen teniéndose.

Ahora estoy en condiciones de definir dos relaciones generales:

Conexividad psicológica es el tener lugar de conexiones psicológicas directas y concretas.

Continuidad psicológica es el tener lugar de cadenas parcialmente superpuestas de conexividad *fuerte*.

De estas dos relaciones generales, la conexividad es la más importante tanto en la teoría como en la práctica. La conexividad puede darse en cualquier grado. Entre X hoy e Y ayer podría haber

varios miles de conexiones psicológicas directas, o sólo una conexión. Si hubiera sólo una conexión, X e Y no serían, según la tesis lockeana revisada, la misma persona. Para que X e Y sean la misma persona, tiene que haber durante cada día *las suficientes* conexiones psicológicas directas. Como la conexividad es una cuestión de grado, no podemos definir convincentemente con precisión qué cuenta como suficiente. Pero podemos decir que hay conexividad suficiente si el número de conexiones directas, durante cada día, es *como mínimo la mitad* del número que se da, durante cada día, en la vida de casi toda persona real [6]. Cuando hay suficientes conexiones directas, tenemos lo que llamo conexividad *fuerte*.

¿Podríamos tener en esta relación el criterio de identidad personal? Una relación *F* es *transitiva* si se cumple que, si X está F-relacionada con Y, e Y está F-relacionada con Z, X y Z *tienen* que estar F-relacionadas. La identidad personal es una relación transitiva. Si Bertie fue la misma persona que el filósofo Russell, y Russell fue la misma persona que el autor de *Por qué no soy cristiano*, este autor y Bertie tienen que ser la misma persona.

La conexividad fuerte *no* es una relación transitiva. Ahora yo estoy fuertemente conectado a mí mismo ayer, cuando estaba fuertemente conectado a mí mismo hace dos días, cuando estaba fuertemente conectado a mí mismo hace tres días, y así sucesivamente. Pero no se sigue que esté ahora fuertemente conectado a mí mismo hace veinte años, cosa que desde luego no ocurre. Entre yo ahora y yo mismo hace veinte años hay muchas menos conexiones psicológicas directas que las que se dan durante cualquier día en la vida de casi todos los adultos. Por ejemplo, mientras que la mayor parte de los adultos tienen muchos recuerdos de experiencias que tuvieron el día anterior, yo tengo pocos recuerdos de experiencias que tuve cualquier día de hace veinte años.

[6] Esta sugerencia necesitaría ampliación, puesto que son muchos los modos de contabilizar el número de conexiones directas. Y a ciertas clases de conexión se les debería dar más importancia que a otras. Como sugiero más adelante, se le debería dar mayor peso a las conexiones que son distintivas, o diferentes en diferentes personas. (Todos los angloparlantes, por ejemplo, comparten muchos recuerdos no distintivos de cómo hablar inglés.)

Por «el criterio de identidad personal a través del tiempo» me refiero a lo que esta identidad *necesariamente implica o a aquello en lo que consiste*. Como la identidad es una relación transitiva, el criterio de identidad tiene que ser también una relación transitiva. Como la conexividad fuerte no es transitiva, no puede ser el criterio de identidad. Y acabo de describir un caso en el que esto es claro. Soy la misma persona que yo mismo hace veinte años, aunque ahora no esté fuertemente conectado a mí mismo entonces.

Aunque un defensor de la concepción de Locke no puede apelar a la conexividad psicológica, sí puede apelar a la continuidad psicológica, que *es* transitiva. Puede apelar a

> *El criterio psicológico*: (1) Hay *continuidad psicológica* si y sólo si hay cadenas parcialmente superpuestas de conexividad fuerte. X hoy es la misma persona que Y en algún momento pasado si y sólo si (2) X es psicológicamente continuo con Y, (3) esta continuidad tiene la clase correcta de causa, y (4) no ha tomado una forma «ramificada». (5) La identidad personal a través del tiempo consiste justamente en el darse hechos como los que van de (2) a (4).

Como con el criterio físico, (4) lo explicaremos después.

Hay tres versiones del criterio psicológico. Difieren en la cuestión de cuál es la clase *correcta* de causa. Según la versión *restringida*, tiene que ser la causa *normal*. Según la versión *amplia*, podría ser *cualquier* causa *fiable*. Según la versión *amplísima*, la causa podría ser *cualquiera*.

El criterio psicológico restringido emplea las palabras en su sentido corriente. De modo que yo recuerdo haber tenido una experiencia sólo si

> (1) parece que recuerdo haber tenido una experiencia,
>
> (2) tuve esa experiencia,

y

> (3) mi recuerdo aparente es causalmente dependiente, del modo normal, de esta experiencia pasada

Que necesitamos la condición (3) lo podemos poner de manifiesto con un ejemplo. Supongamos que perdí el conocimiento dándome un golpe en un accidente de escalada. Una vez que me recupero, el escalador que me acompañaba me habla del grito que dio justo antes de que yo me cayera. Y podría ser cierto que yo haya tenido la experiencia de ese grito. Pero aunque se cumplen las condiciones (1) y (2), no debemos pensar que estoy recordando esa experiencia pasada. Es un hecho bien establecido el que la gente nunca puede recordar sus últimas experiencias antes de quedarse sin conocimiento por un golpe en la cabeza. Por eso debemos decir que mi recuerdo aparente de haber oído gritar a mi compañero no es un recuerdo real de esa experiencia pasada. Tengo este recuerdo aparente sólo porque mi compañero me contó después que él había gritado [7].

Observaciones similares se aplican a las otras clases de continuidad, como la continuidad de carácter. Según el criterio psicológico restringido, aunque el carácter de una persona cambie radicalmente, hay continuidad de carácter si estos cambios tienen una de las diversas causas normales posibles. Algunos cambios de carácter se llevan a cabo deliberadamente; otros son la consecuencia natural de la edad; otros la respuesta natural a determinadas experiencias. Pero no habría continuidad de carácter si se produjeran cambios radicales y no deseados debido a interferencias anormales, como por ejemplo la manipulación directa del cerebro.

Aunque es la memoria la que nos hace conscientes de nuestra propia existencia continua a través del tiempo, las otras continuidades tienen gran importancia. Podemos pensar incluso que tienen la suficiente importancia como para proporcionar identidad personal aun en ausencia de la memoria. Por eso defenderemos lo que Locke negó, que una persona continúa existiendo aunque padezca una amnesia completa.

Además de la versión restringida, describí las dos versiones amplias del criterio psicológico. Estas versiones amplían el sentido de varias palabras. Según el sentido corriente de «recuerdo», un

[7] Sigo a Martin y Deutscher.

recuerdo tiene que tener su causa normal. Los dos criterios psicológicos amplios apelan a un sentido amplio de «recuerdo», que permite o una causa fiable o cualquier causa. Afirmaciones similares se aplican a las otras clases de conexión psicológica directa. Para simplificar mi discusión de estos tres criterios, usaré «continuidad psicológica» en su sentido más amplio, que permite que esta continuidad tenga *cualquier* causa.

Si apelamos a la versión restringida, que insiste en la causa normal, el criterio psicológico coincide en la mayor parte de los casos con el criterio físico. Las causas normales del recuerdo conllevan la existencia continua del cerebro. Y una parte o la totalidad de nuestros rasgos psicológicos dependen de estados o sucesos de nuestros cerebros. La existencia continua del cerebro de una persona es, como mínimo, parte de la causa normal de la continuidad psicológica. Según el criterio físico, una persona sigue existiendo si y sólo si *(a)* sigue existiendo *suficiente* cerebro de esta persona como para que continúe siendo el cerebro de una persona viva, y *(b)* no ha habido ramificación en esta continuidad física. Se defiende que *(a)* y *(b)* con las condiciones necesarias y suficientes para la identidad de esta persona, o para su existencia continua, a través del tiempo. Según el criterio psicológico restringido, *(a)* es necesario, pero no suficiente. Una persona sigue existiendo si y sólo si *(c)* hay continuidad psicológica, *(d)* esta continuidad tiene su causa normal, y *(e)* no ha tomado una forma ramificada. Se exige *(a)* como parte de la causa normal de la continuidad psicológica.

Reconsideremos el comienzo de mi historia imaginaria, cuando se destruían mi cerebro y mi cuerpo. El escáner y el replicador producen una persona que tiene un nuevo cerebro y un nuevo cuerpo, pero exactamente iguales, y que es psicológicamente continua conmigo tal y como yo era cuando apreté el botón verde. La causa de esta continuidad es, aunque inusual, fiable. Tanto según el criterio físico como según el criterio psicológico restringido, mi Réplica *no* sería yo. Según el criterio amplio, *sería* yo.

Defenderé que no necesitamos decidir entre estas tres versiones del criterio psicológico. Una analogía parcial puede sugerir por qué. Supongamos que unas personas se están quedando ciegas por tener

los ojos dañados. Los científicos están desarrollando ahora unos ojos artificiales, que consisten en lentes de cristal o de plástico, y un microordenador que envía a través del nervio óptico señales eléctricas similares a las que envía a través de ese nervio un ojo natural. Cuando estos ojos artificiales estén más avanzados, podrían proporcionar a los que se hayan quedado ciegos experiencias visuales justo como las que solían tener. De forma que lo que a una de estas personas le dé la sensación de ver, correspondería a lo que se halla en efecto ante ella. Y sus experiencias visuales serían causalmente dependientes, de esta manera nueva pero fiable, de las ondas luminosas que vienen de los objetos que están delante.

¿Estaría *viendo* la persona esos objetos? Si insistimos en que ver tiene que implicar la causa normal, responderíamos No. Pero aunque la persona no pueda ver, lo que tiene es *tan bueno como* ver, no sólo como un modo de saber lo que está dentro del radio de acción de la vista, sino también como una fuente de placer visual. Si aceptáramos el criterio psicológico, podríamos hacer una afirmación parecida. Si la continuidad psicológica no tiene su causa normal, puede que no proporcione identidad personal, pero podemos afirmar que, aun así, lo que proporciona es *tan bueno como* la identidad personal.

79. Las otras concepciones

Estoy inquiriendo cuál es el criterio de la identidad personal a través del tiempo —qué implica esta identidad, en qué consiste—. En primer lugar describí la continuidad física espacio-temporal que, según la concepción estándar, sería el criterio de identidad de los objetos físicos. Luego describí dos concepciones de la identidad personal, los criterios físico y psicológico.

Hay una asunción sobre estas concepciones que resulta natural pero falsa. Muchos creen en lo que se llama *Materialismo* o *Fisicalismo*. Esta es la tesis de que no hay objetos, estados ni sucesos puramente mentales. Según una versión del Fisicalismo, todo suceso mental no es más que un suceso físico en un cerebro y un sistema nervioso particular. Hay otras versiones. Los que no son fisicalistas son o *dualistas* o *idealistas*. Los dualistas creen que los sucesos mentales *no* son sucesos físicos, lo que puede ser cierto por mucho que todos los sucesos mentales dependan causalmente de sucesos físicos en un cerebro. Los idealistas creen que todos los estados y sucesos son, cuando los comprendemos correctamente, puramente mentales. Dadas estas distinciones, puede que demos por sentado que los fisicalistas tienen que aceptar el criterio físico de identidad personal.

Pero no es así: los fisicalistas podrían aceptar el criterio psicológico, y hasta podrían aceptar la versión que permite una causa fiable, o cualquier causa. Podrían por lo tanto pensar que, en el teletransporte simple, mi Réplica sería yo, con lo que estarían rechazando aquí el criterio físico [8].

Estos criterios no son las únicas concepciones de la identidad personal. Ahora describiré algunas de las otras concepciones, las que o son lo suficientemente convincentes o tienen numerosos partidarios, por lo que vale la pena que las consideremos. Puede que sea difícil seguir esta descripción, pero dará una idea aproximada de lo que nos espera más adelante. Si buena parte de este resumen parece oscuro o trivial, no hay que preocuparse.

Empiezo con una nueva distinción. Según el criterio físico, la identidad personal a través del tiempo implica sólo la existencia físicamente continua de suficiente cerebro como para que siga siendo el cerebro de una persona viva. Según el criterio psicológico, la identidad personal a través del tiempo implica sencillamente las diversas clases de continuidad psicológica, con el tipo correcto de causa. Estas dos concepciones son *reduccionistas* porque establecen

(1) que el hecho de la identidad de una persona a través del tiempo consiste sólo en el darse de determinados hechos más concretos.

[8] Quinton defiende esta concepción.

También pueden afirmar

> (2) que estos hechos pueden describirse sin presuponer la identidad de la persona en cuestión, ni afirmar explícitamente que las experiencias de la vida de esta persona son tenidas por la persona en cuestión, ni tampoco afirmar explícitamente que la persona en cuestión existe. Estos hechos se pueden describir de un modo *impersonal*.

A lo mejor parece que (2) podría no ser verdadero. Cuando describimos la continuidad psicológica que unifica la vida mental de una persona, tenemos que mencionar a esta persona, y a otras muchas, al describir el *contenido* de muchos pensamientos, deseos, intenciones, y otros estados mentales. Pero mencionar a esta persona de este modo no implica ni afirmar que estos estados mentales son tenidos por la persona en cuestión, ni tampoco afirmar que esta persona existe. Estas afirmaciones necesitan argumentos adicionales, que más tarde daré.

Nuestra concepción será *no reduccionista* si rechazamos las dos tesis reduccionistas.

Muchos no reduccionistas creen que *somos entidades que existen separadamente*. Según esta concepción, la identidad personal a través del tiempo no sólo consiste en continuidad física y/o psicológica, sino que implica un hecho adicional. Una persona es una entidad que existe de forma separada, distinta de su cerebro y de su cuerpo, y de sus experiencias. Según la versión mejor conocida de esta idea, una persona es una entidad *puramente mental*: un Ego Puro Cartesiano, una sustancia espiritual. Pero también se podría pensar que una persona es una entidad *física* que existe de forma separada, una entidad de una clase que todavía no se ha reconocido en las teorías de la física contemporánea.

Hay otra concepción no reduccionista, la que niega que seamos entidades que existen separadamente, distintas de nuestros cerebros y nuestros cuerpos, y de nuestras experiencias, pero establece que, aunque no somos entidades que existan separadamente, la identidad personal *es* un hecho adicional, que no consiste sólo en la continuidad física y/o psicológica. La llamo *tesis del hecho adicional*.

Trazaré ahora algunas distinciones. Los criterios físico y psicológico son versiones de la concepción reduccionista; y hay diferentes versiones de cada criterio. Pero lo que está necesariamente implicado en la existencia continua de una persona es menos que lo que está de hecho implicado. De forma que mientras que los que creen en los diferentes criterios están en desacuerdo en lo que respecta a los casos imaginarios, están de acuerdo sobre lo que está de hecho implicado en la existencia continua de la mayoría de las personas reales. Empezarían a no estar de acuerdo únicamente si, por ejemplo, la gente comenzase a ser teletransportada.

Según la concepción reduccionista, la existencia de cada persona no implica otra cosa que la existencia de un cerebro y un cuerpo, la realización de determinados actos, el pensar ciertos pensamientos, la ocurrencia de determinadas experiencias, y así sucesivamente. Ayudará ampliar el sentido corriente de la palabra «suceso». Usaré esta palabra para incluir hasta sucesos tan *aburridos* como la existencia continua de una creencia o de un deseo. Este uso hace a la concepción reduccionista más fácil de describir. Y elude lo que pienso que es una implicación engañosa de las palabras «estado mental». Mientras que un estado tiene que ser un estado *de* una entidad, esto no ocurre con un suceso. Dado este uso ampliado de la palabra «suceso», todos los reduccionistas aceptarían

> (3) La existencia de una persona no consiste en otra cosa que en la existencia de un cerebro y de un cuerpo, y en la ocurrencia de una serie de sucesos físicos y mentales interrelacionados.

Algunos reduccionistas mantienen

> (4) Una persona *no es más que* un cerebro y un cuerpo concretos, y una serie de sucesos interrelacionados.

Otros reduccionistas afirman

> (5) Una persona es una entidad que es *distinta* de un cerebro y un cuerpo, y de una serie de sucesos.

Según esta versión de la tesis reduccionista, una persona no es meramente un objeto compuesto, con varios componentes determinados. Una persona es una entidad que *tiene* un cerebro y un cuerpo, y que *tiene* pensamientos y deseos concretos, y así sucesivamente. Pero, aunque (5) sea verdadera, una persona no es una entidad *que exista separadamente*. Aunque (5) sea verdadera, (3) también lo es.

Esta versión del reduccionismo puede parecer internamente contradictoria. Quizás (3) y (5) parezcan inconsistentes. Puede ayudar que consideremos la analogía de Hume: «No puedo comparar el alma con nada mejor que con una república o comunidad política» [9]. La mayor parte de nosotros es reduccionista en lo que hace a las naciones. Aceptaríamos afirmaciones como éstas: las naciones existen. Ruritania no existe, pero Francia sí. Aunque las naciones existen, una nación no es una entidad que exista de forma separada, aparte de sus ciudadanos y de su territorio. Nosotros aceptaríamos

(6) La existencia de una nación implica sólo la existencia de sus ciudadanos, viviendo juntos de determinadas maneras en su territorio.

Unos afirman

(7) Una nación no *es* otra cosa que esos ciudadanos y ese territorio.

Otros sostienen

(8) Una nación es una entidad distinta de sus ciudadanos y de su territorio.

Por las razones que planteo en el Apéndice D, podemos pensar que (6) y (8) no son inconsistentes. Si lo pensamos así, podemos aceptar que no hay inconsistencia entre las afirmaciones correspondientes (3) y (5). Así que podemos estar de acuerdo en que la versión del Reduccionismo expresada en (3) y en (5) es una concepción consistente. Si esta versión es consistente, como pienso, es la mejor versión. Está cerca de nuestro concepto real de persona. Pero en la mayor parte de lo que viene a continuación podemos ignorar la diferencia entre estas dos versiones [10].

[9] Hume (1), Parte IV, Sección 6, reeditado en Perry (1).

Además de afirmar (1) y (2), los reduccionistas podrían afirmar también

(9) Aunque las personas existen, podríamos dar una descripción *completa* de la realidad *sin* afirmar que existan personas.

Llamo a esto la tesis de *que una descripción completa podría ser impersonal*.

Esta tesis también puede parecer internamente contradictoria. Si existen las personas, y una descripción de lo que existe no menciona a las personas, ¿cómo puede ser completa esta descripción?

Un reduccionista podría dar la siguiente respuesta. Supongamos que un objeto tiene dos nombres. Esto ocurre con el planeta que se llama *Venus* y la *Estrella de la Tarde*. En nuestra descripción de lo que existe, podríamos afirmar que existe Venus. Nuestra descripción podría entonces ser completa aunque no afirmáramos que la Estrella de la Tarde existe. No tendríamos necesidad de hacer esta afirmación ya que, usando su otro nombre, ya hemos afirmado que este objeto existe.

Una afirmación similar es de aplicación cuando un hecho puede describirse de dos modos. Algunos reduccionistas aceptan (4), la afirmación de que una persona es un cerebro y un cuerpo concretos, y una serie de sucesos físicos y mentales interrelacionados. Si esto es lo que *es* una persona, podemos describir este hecho diciendo o bien

(10) que existen un cerebro y un cuerpo concretos, y una serie concreta de sucesos físicos y mentales interrelacionados.

[10] Secciones 96 y 98-9, y Capítulo 14.

O bien

(11) que existe una persona concreta.

Si (10) y (11) son dos modos de describir el *mismo* hecho, una descripción, para ser completa, no tiene necesidad de hacer *ambas* afirmaciones. Bastaría con hacer la afirmación (10). Aunque esta persona exista, una descripción completa no necesita afirmar que existe, puesto que de este hecho ya se ha informado en la afirmación (10).

Otros reduccionistas aceptan (5), la afirmación de que una persona es distinta de su cerebro y de su cuerpo, y de sus actos, pensamientos, y otros sucesos físicos y mentales. Según esta versión del reduccionismo, la afirmación (10) no describe el mismísimo hecho que describe la afirmación (11). Pero la afirmación (10) puede *implicar* la (11). Para ponerlo más comedidamente: dada nuestra comprensión del concepto de persona, si sabemos que (10) es verdadera, sabemos que (11) es verdadera. Estos reduccionistas pueden decir que, si nuestra descripción de la realidad formula o implica, o nos permite conocer, la existencia de todo lo que existe, nuestra descripción es completa. Esta afirmación no es tan obviamente verdadera como la afirmación de que una descripción completa no necesita dar dos descripciones del mismo hecho. Pero parece plausible. Si está justificada, y la concepción reduccionista es verdadera, estos reduccionistas pueden describir la realidad de forma completa sin afirmar que las personas existen [11].

[11] El Reduccionismo levanta difíciles cuestiones, como es bien sabido. Estoy influido por estas observaciones en Kripke, p. 271:

«Aunque el enunciado de que Inglaterra luchó contra Alemania en 1943 tal vez no pueda ser *reducido* a ningún enunciado sobre individuos, sin embargo en cierto sentido no es un hecho "por encima y además de" la colección de todos los hechos sobre personas, y su conducta a lo largo de la historia. El sentido en que los hechos sobre las naciones no son hechos "por encima y además de" los que versan sobre personas puede expresarse en la observación de que una descripción del mundo que mencione todos los hechos sobre las personas pero que omita los hechos sobre las naciones puede ser una descripción *completa* del mundo, de la que se siguen los hechos sobre las naciones. De forma similar, tal vez los hechos sobre objetos mate-

Mis afirmaciones sobre el reduccionismo trazan distinciones que, en esta forma abstracta, son difíciles de captar. Pero hay otras formas de descubrir si somos reduccionistas en nuestra opinión sobre cosas de diversos tipos. Si aceptamos una concepción reduccionista, pensaremos que la identidad de estas cosas puede ser, de un modo en absoluto desconcertante, *indeterminada*. Si *no* pensamos así, probablemente seamos no-reduccionistas acerca de estas clases de cosas.

Consideremos, por ejemplo, los clubes. Supongamos que cierto club existe durante varios años, reuniéndose con regularidad. De repente se interrumpen las reuniones. Unos años después, algunos de los miembros del club forman un club con el mismo nombre y las mismas reglas. Preguntamos: «¿Han vuelto a convocar estas personas al *mismo* club? ¿O meramente han fundado *otro* club, que es exactamente igual?». Podría haber una respuesta a esta pregunta. El club original podría haber contado con una regla que explicase cómo, tras un período de no existencia, podría ser reconvocado. O podría haber tenido una regla que lo impidiese. Pero supongamos

riales no son hechos "por encima y además de" los hechos sobre sus moléculas constituyentes. Entonces podemos preguntar, dada una descripción de una situación posible no actualizada en términos de personas, si todavía existe Inglaterra en esa situación... De forma similar, dadas ciertas vicisitudes contrafácticas en la historia de las moléculas de una mesa, *M*, uno puede preguntar si *M* existiría en esa situación, o si un determinado manojo de moléculas, que en esa situación constituiría una mesa, constituye la misma mesa *M*. En cada caso, pedimos criterios de identidad a través de mundos posibles para determinados particulares en términos de los criterios para otros particulares más "básicos". Si los enunciados acerca de naciones (o tribus [mesas?]) no son *reducibles* a aquellos acerca de otros constituyentes más "básicos", si hay una "textura abierta" en la relación entre ellos, casi no podemos esperar dar criterios de identidad sólidos y firmes. Sin embargo, en casos concretos podemos ser capaces de contestar si un determinado manojo de moléculas todavía constituiría la mesa *M*, aunque en algunos casos la respuesta puede ser indeterminada. Pienso que consideraciones similares se aplican al problema de la identidad a través del tiempo...»

Dada la posible no reducibilidad del enunciado sobre Inglaterra, me inclino por debilitar la expresión "se siguen" al final de la segunda oración de Kripke. La cuestión central sobre la identidad personal creo que es la de si estas observaciones se aplican también a las personas, y no sólo a las naciones y a las mesas.

que no hay tal regla, ni tampoco hechos legales, que apoyen ninguna de las dos respuestas a nuestra pregunta. Y supongamos que la gente involucrada, si se les planteara nuestra pregunta, no le diera una respuesta. No habría entonces respuesta a nuestra pregunta. La afirmación «Este es el mismo club» no sería *ni verdadera ni falsa*.

Aunque no hay respuesta a nuestra pregunta, puede que no haya nada que no sepamos. Y es que la existencia de un club no está separada de la existencia de sus miembros, actuando juntos de ciertos modos. La existencia continua de un club sólo conlleva que sus miembros mantengan encuentros, encuentros que se conduzcan según las reglas del club. Si sabemos todos los hechos sobre cómo mantuvo la gente sus encuentros, y sobre las reglas del club, sabemos todo lo que hay que saber. Por eso no nos quedaríamos perplejos al no poder responder a la pregunta, «¿Es éste el mismo club?». No nos quedaríamos perplejos porque, aun sin contestar a esta pregunta, podemos saber todo lo que sucedió. Si esto ocurre con una pregunta, diré que esta pregunta es *vacía*.

Cuando hacemos una pregunta vacía, estamos considerando sólo un hecho o resultado. Las diferentes respuestas a nuestra pregunta nada más que son diferentes descripciones de este hecho o resultado. Por eso podemos saber todo lo que hay que saber sin responder a esta pregunta vacía. En el ejemplo que puse podemos preguntar, «¿Es el mismo club o simplemente otro club exactamente igual?». Pero no se trata con esto de dos posibilidades diferentes, de las cuales una tenga que ser verdadera.

Cuando una pregunta vacía no tiene respuesta, podemos decidir *dar*le una respuesta. Podríamos decidir llamar al último club el mismo que el original. O podríamos decidir decir que es otro club, exactamente igual. Esta no es una decisión entre diferentes concepciones de lo que realmente ocurrió. Antes de tomar nuestra decisión ya sabíamos lo que ocurrió. Simplemente elegimos una de dos descripciones diferentes del mismo curso de sucesos.

Si somos reduccionistas en lo relativo a la identidad personal, deberíamos hacer afirmaciones parecidas. Podemos describir casos en que, entre yo ahora y una persona futura, las conexiones físicas y psicológicas se mantienen sólo en un grado reducido. Si me imagino a mí mismo en un caso así, siempre puedo preguntar, «¿Estoy a punto de morir? ¿La persona resultante será yo?». Según la concepción reduccionista, en ciertos casos no habría respuesta. Mi pregunta sería *vacía*. La afirmación de que estaba a punto de morir no sería ni verdadera ni falsa. Si conociera los hechos acerca de la continuidad física y la conexividad psicológica, sabría todo lo que habría que saber. Lo sabría todo, aunque no supiera si estaba a punto de morir o seguiría viviendo durante muchos años.

Cuando se nos aplica a nosotros mismos, esta tesis reduccionista resulta difícil de creer. En casos imaginarios de esta clase, algo no usual está a punto de ocurrir. Pero la mayoría de nosotros se inclina a creer que, en cualquier caso concebible, la pregunta «¿Estoy a punto de morir?» tiene que tener una respuesta. Y nos inclinamos a pensar que esta respuesta tiene que ser, de la manera más simple, un Sí o un No. Cualquier persona futura tiene que ser yo o alguien diferente. Llamo a estas creencias la tesis de que *nuestra identidad tiene que ser determinada*.

A continuación describiré dos tesis explicativas. La primera responde a una nueva pregunta. ¿Qué da unidad a las diferentes experiencias que son tenidas por una persona singular en el mismo momento? Mientras escribo esta frase, soy consciente del movimiento de mis dedos, puedo ver la luz del sol sobre mi mesa de trabajo y puedo oír el viento agitando las hojas de los árboles. ¿Qué unifica estas diferentes experiencias? Son las experiencias que están siendo tenidas, en este preciso momento, por una persona concreta, un *sujeto de experiencias*. Y una pregunta parecida incluye la totalidad de mi vida. ¿Qué da unidad a las diferentes experiencias que, juntas, constituyen esta vida? Hay quienes dan la misma respuesta. Lo que unifica todas estas experiencias es, simplemente, que son todas mías. A estas respuestas las denomino la tesis de que *la unidad psicológica se explica por la propiedad*.

Las tesis descritas hasta aquí tratan sobre la naturaleza de la identidad personal. Terminaré con un par de tesis que tratan, no de la naturaleza de esta identidad sino de su importancia. Consi-

deremos un caso corriente en que, incluso según cualquier versión de la concepción reduccionista, hay dos resultados posibles. En uno de los resultados, estoy a punto de morir. En el otro, viviré muchos años. Si estos años valiesen la pena vivirse, el segundo resultado sería mejor para mí. Y la diferencia entre los dos resultados la juzgaríamos importante, según la mayoría de las teorías de la racionalidad y la mayoría de las teorías morales. Tendría una relevancia racional y moral el que yo estuviera a punto de morir o, por el contrario, fuera a vivir muchos años. Lo que se juzga importante aquí es si, durante esos años, habrá alguien vivo que *será yo*. Esta es una cuestión sobre la identidad personal. Según cierta opinión, en esta clase de casos esto es siempre lo importante. Llamo a esto la tesis de que *la identidad personal es lo que importa*. Esta es la concepción natural.

La tesis rival es que *la identidad personal no es lo que importa*. Yo defiendo que

Lo que importa es la relación R: conexividad y/o continuidad psicológica, con la clase correcta de causa.

Ya que es más controvertido, añado, como afirmación separada

En una explicación de lo que importa, la clase correcta de causa podría ser cualquier causa.

Es en los casos imaginarios donde mejor podemos decidir si lo que importa es la relación R o la identidad personal. Un ejemplo puede ser el caso de la línea secundaria, en el que mi vida se solapa brevemente con la de mi Réplica. Supongamos que pensamos que yo y mi Réplica somos dos personas diferentes. Yo estoy a punto de morir, pero mi Réplica vivirá otros cuarenta años. Si la identidad personal es lo que importa, yo debería considerar mi perspectiva aquí claramente tan mala como la muerte normal. Pero si lo que importa es la relación R, con cualquier causa, yo debería considerar este modo de morir casi tan bueno como la supervivencia normal.

El desacuerdo entre estas concepciones no se limita a los casos imaginarios. Las dos tesis también discrepan sobre todas las vidas reales que se viven. El desacuerdo es aquí menos pronunciado, puesto que, según ambas tesis, todas estas vidas, o casi todas, contienen la relación que importa. Según todas las concepciones plausibles de la naturaleza de la identidad personal, ésta casi siempre coincide con

la continuidad psicológica, y coincide aproximadamente con la conexividad psicológica. Pero, como mantendré después, supone una gran diferencia cuál de estas pensamos que es la que importa. Si dejamos de creer que nuestra identidad es lo que importa, esto puede afectar a algunas de nuestras emociones, en especial las que están relacionadas con nuestra actitud ante el envejecimiento y la muerte. Y, como defenderé, puede que cambiemos nuestras ideas sobre la racionalidad y la moralidad.

Ahora he dado una descripción preliminar de diversas concepciones diferentes. Formuladas de esta forma abstracta, puede que la descripción no pueda ser completamente clara. Pero cuando las someta a discusión tal vez se torne claro lo que ahora está oscuro.

¿Cómo se relacionan entre sí estas concepciones? Afirmaré lo que algunos niegan, que muchas de ellas se mantienen o caen juntas. Si esto es cierto, será más fácil decidir cuál es la verdad. Cuando veamos cómo se relacionan estas concepciones, encontraremos, eso creo, que sólo tenemos dos alternativas. Vale la pena formular con antelación algunos de los modos en que, como defenderé, se relacionan estas concepciones.

Si no creemos que somos entidades que existen separadamente, ¿podemos pensar justificadamente que la identidad personal es lo que importa? Algunos creen que podemos. Yo sostendré que no podemos.

Si no pensamos que somos entidades que existen separadamente, ¿podemos pensar justificadamente que la identidad personal no consiste sólo en continuidad física y psicológica, sino que es un hecho adicional? Pienso que no podemos.

Si pensamos que nuestra identidad tiene que ser determinada, ¿tenemos que creer que somos entidades que existen separadamente? Tener la primera creencia no implica tener la segunda. Podríamos pensar tanto que no somos entidades que existen separadamente como que, para cualquier pregunta sobre la identidad personal, tiene que haber siempre una respuesta, que tiene que ser Sí o No. Los hay que aceptan este modo de ver las cosas. Pero yo defenderé que es injustificable. Sólo si somos entidades que existen sepa-

radamente puede ser cierto que nuestra identidad tiene que ser determinada.

Sería posible afirmar que somos entidades que existen separadamente, pero negar que nuestra identidad tenga que ser determinada. Pero hay pocos que combinarían estas dos afirmaciones.

Supongamos a continuación que pensamos que la unidad psicológica se explica por la propiedad. Pensamos que la unidad de conciencia de una persona en un momento dado se explica por el hecho de que las diferentes experiencias de esa persona están siendo todas tenidas por ella. Y pensamos que la unidad de la vida completa de una persona se explica por el hecho de que todas las experiencias de esta vida son tenidas por ella. Estas son las explicaciones que dan los que afirman que somos entidades que existen separadamente. ¿Podemos dar estas explicaciones si rechazamos esa afirmación? Algunos sugieren que sí, yo defenderé que no.

También defenderé las siguientes conclusiones:

(1) No somos entidades que existan separadamente, aparte de nuestros cerebros y de nuestros cuerpos, y de diversos sucesos físicos y mentales interrelacionados. Nuestra existencia conlleva sencillamente la existencia de nuestros cerebros y de nuestros cuerpos, y el llevar a cabo nuestros actos, y el pensar nuestros pensamientos, y la ocurrencia de otros sucesos físicos y mentales determinados. Nuestra identidad a través del tiempo no implica otra cosa que *(a)* la relación R —conexividad psicológica y/o continuidad psicológica— con la clase correcta de causa, dando por hecho *(b)* que esta relación no adopta una forma «ramificada», dándose entre una persona y dos personas futuras diferentes.

(2) No es cierto que nuestra identidad sea siempre determinada. Siempre puedo preguntar, «¿Estoy a punto de morir?». Pero no es cierto que, en todo caso, esta pregunta tenga que tener una respuesta que tenga que ser Sí o No. En algunos casos sería una pregunta vacía.

(3) Hay que explicar dos unidades: la unidad de conciencia en un momento dado, y la unidad de una vida completa. Ninguna de ellas puede explicarse afirmando que diferentes experiencias son tenidas por la misma persona. Tienen que explicarse describiendo las relaciones entre estas numerosas experiencias, y sus relaciones con el cerebro de la persona en cuestión. Y podemos referirnos a estas experiencias, y describir por completo las relaciones que se dan entre ellas, sin afirmar que son tenidas por una persona.

(4) La identidad personal no es lo que importa. Lo que fundamentalmente importa es la relación R, con cualquier causa. Esta relación es lo que importa aunque, en un caso en que una persona esté R-relacionada con otras dos personas, la relación R no proporcione identidad personal. Hay otras dos relaciones que pueden tener alguna importancia: la continuidad física y la similitud física. (Hay personas, las que son muy bellas, para las que la similitud física puede tener gran importancia.)

Aquí va un breve esquema de cómo abogaré por mis conclusiones. Primero trataré de contestar algunas objeciones a mi afirmación de que podríamos describir nuestras vidas de un modo *impersonal*. Luego trataré de demostrar que, aunque no seamos conscientes de ello, estamos naturalmente inclinados a pensar que nuestra identidad tiene que ser siempre determinada. Estamos inclinados a pensar, con toda seguridad, que *tiene* que ser así. A continuación argumentaré que esta creencia natural no puede ser cierta como no seamos entidades que existen separadamente. Entonces argumentaré a favor de la conclusión (1), que dice que no somos entidades de esa clase. Defenderé que, puesto que (1) es verdadera, también lo son mis tres conclusiones restantes.

La mayoría de nosotros aceptaría alguna de las afirmaciones que voy a negar. De modo que argumentaré que la mayoría de nosotros tiene una falsa visión de nosotros mismos y de nuestras vidas reales. Si llegamos a ver que esta visión es falsa, algo puede cambiar en nuestra vida.

11

CÓMO NO SOMOS LO QUE CREEMOS

Las diferentes concepciones de la identidad personal hacen declaraciones diferentes sobre las personas reales y las vidas corrientes. Pero la diferencia que hay entre ellas se hace más clara cuando consideramos ciertos casos imaginarios. La mayoría de los argumentos que discutiré apelan, en parte, a casos de ese tipo. Puede ser imposible que alguno de esos casos ocurra, por muchos progresos que hagan la ciencia y la tecnología. Distingo dos clases de casos. Los hay que van en contra de las leyes de la naturaleza. A estos los llamo *radicalmente* imposibles. Otros no son *nada más que técnicamente* imposibles. ... 401

¿Tiene importancia que un caso imaginario nunca sea posible? Esto depende enteramente de la cuestión que tratemos, o de lo que estemos intentando demostrar. Incluso en la ciencia puede ser útil considerar casos radicalmente imposibles. Un ejemplo es el experimento mental de Einstein en el que preguntó qué vería si pudiera viajar junto a un haz de luz a la velocidad de la luz. Como pone de manifiesto este ejemplo, no es necesario limitarse a considerar sólo casos posibles. Pero no deberíamos pasar por alto que, dependiendo de la cuestión que estemos tratando, la imposibilidad puede hacer irrelevante a un experimento mental.

Comienzo con una objeción al criterio psicológico.

80. ¿La continuidad psicológica presupone la identidad personal?

Recuerdo haber intentado mantenerme en pie, cuando era niño, en medio de las olas del Océano Atlántico estrellándose contra mí. Soy la misma persona que el niño que tuvo esa experiencia. Según la idea de Locke, lo que me hace ser la misma persona que ese niño es mi recuerdo, o «conciencia», de esa experiencia...

El obispo Butler pensaba que esto era un «increíble error». Como escribió, «es en sí mismo evidente que la conciencia de la identidad personal presupone la identidad personal, y por eso no puede constituirla, del mismo modo que el conocimiento, en cualquier caso, no puede constituir la verdad, puesto que la presupone» [12].

Ya he revisado la concepción de Locke. El criterio psicológico apela, no a recuerdos singulares, sino a la continuidad de la memoria, y, más generalmente, a la relación R, que incluye otros tipos de continuidad psicológica. Pero esta revisión no contesta a la objeción de Butler.

Según una interpretación, la objeción sería esta: «Forma parte de nuestro concepto de memoria el que podamos recordar sólo *nuestras propias* experiencias. La continuidad de memoria presupone por consiguiente la identidad personal. Y lo mismo ocurre por tanto con tu relación R. Afirmas que la identidad personal consiste simplemente en el darse de la relación R. Pero esto debe ser falso si la misma relación R presupone la identidad personal».

Para responder a esta objeción, podemos definir un concepto más amplio, el de *cuasimemoria*. Tengo un cuasi-recuerdo fiel de una experiencia pasada si

(1) parezco recordar tener una experiencia,

[12] Butler, p. 100.

(2) *alguien* tuvo esta experiencia

y

(3) mi recuerdo aparente es causalmente dependiente, del modo correcto, de esa experiencia pasada.

Según esta definición, los recuerdos corrientes son una subclase de los cuasi-recuerdos. Son cuasi-recuerdos de nuestras propias experiencias pasadas [13].

Nosotros no cuasi-recordamos las experiencias pasadas de otras personas. Pero podríamos empezar a hacerlo. Las causas de los recuerdos de larga duración son huellas mnémicas. Antes se pensaba que estas podrían estar localizadas, implicando cambios en sólo unas pocas células cerebrales. Ahora es más probable que una huella mnémica concreta conlleve cambios en un número de células mayor. Supongamos que, aunque esto sea cierto, los neurocirujanos desarrollaran formas de crear en un cerebro una copia de una huella mnémica de otro cerebro. Esto podría permitirnos cuasi-recordar las experiencias pasadas de otras personas.

Consideremos

> *Los Recuerdos Venecianos*. Jane ha accedido a que le copien en su cerebro algunas de las huellas mnémicas de Paul. Tras recobrar la conciencia en la sala post-quirúrgica, se encuentra con que tiene un nuevo conjunto de vívidos recuerdos aparentes. Parece acordarse de caminar sobre el pavimento marmóreo de una plaza, escuchando el aleteo de palomas que vuelan y los chillidos de las gaviotas, y viendo los destellos de luz en el agua verde. Un recuerdo aparente es especialmente claro. Parece acordarse de divisar al otro lado del agua una isla, una isla en la que destaca una blanca iglesia paladiana, recortándose resplandeciente contra una oscura nube de tormenta.

¿Qué es lo que debería pensar Jane de estos recuerdos aparentes? Supongamos que, como ha visto esa iglesia en fotografías, sabe

[13] Sigo a Shoemaker (2).

que se trata de San Giorgio, en Venecia. También sabe que ella nunca ha estado en Italia, mientras que Paul va a Venecia a menudo. Puesto que sabe que ha recibido copias de algunas de las huellas mnémicas de Paul, podría asumir con fundamento que puede estar cuasi-recordando algunas de las experiencias de Paul en Venecia.

Añadamos un detalle más al caso. Jane parece acordarse de haber visto algo extraordinario: el resplandor de un rayo hendido que cayó desde la nube oscura sobre la torre del reloj de San Giorgio, alcanzando también la roja chimenea de un remolcador que pasaba por allí. Le pregunta a Paul si él recuerda haber visto un suceso tan extraordinario. Sí que lo recuerda, y ha guardado el número del *Gazzetino* donde se informa de lo ocurrido. Sabiendo todo esto, Jane no debería descartar su recuerdo aparente como una ilusión. Debe rendirse a la evidencia de que tiene un cuasi-recuerdo fiel de la impresión que le hizo a Paul el resplandor del rayo.

Para que los cuasi-recuerdos de Jane le den conocimiento de las experiencias de Paul, ella tiene que saber más o menos cómo fueron causados. Cosa que no se exige en el caso de los recuerdos corrientes. Quitando esta diferencia, los cuasi-recuerdos proporcionarían una clase similar de conocimiento de las vidas pasadas de otras personas. Proporcionarían conocimiento de cómo fueron estas vidas, *desde dentro*. Cuando Jane parece acordarse de caminar por la Piazza, oyendo las gaviotas y viendo la iglesia blanca, sabe parcialmente cómo fue ser Paul aquel día en Venecia.

En un sentido, los recuerdos aparentes de Jane pueden estar equivocados. Puede decirse: «Como Jane parece acordarse de *haber visto* el rayo, parece acordarse de que *ella misma* vio el rayo. Su recuerdo aparente puede decirle con fidelidad cómo fue la experiencia de Paul, pero le dice, falsamente, que fue *ella* la que tuvo esa experiencia».

Puede haber un sentido en que esta afirmación sea cierta. Los recuerdos aparentes de Jane pueden venirle a ella en lo que Peacocke llama *el modo de presentación de la primera persona* [14]. De modo que cuando parece acordarse de pasear por la Piazza, podría parecer estarse acordando de ver a un niño que corre *hacia ella*. Si esto es lo

[14] En Peacocke (2).

que parece recordar, tiene que parecer acordarse de *ella misma* viendo al niño correr hacia ella.

Podríamos negar estas afirmaciones. En un sueño me puede parecer verme a mí mismo desde un punto de vista *exterior* a mi propio cuerpo. Podría parecerme que me veo a mí mismo venir corriendo hacia este punto de vista. Como soy *yo mismo* el que me parece ver corriendo en esta dirección, esta dirección no puede ser hacia *mí mismo*. Podría decir que me parece verme a mí mismo corriendo hacia *el punto de vista del observador*. Y podría decirse que esta es la dirección en la que Jane parece acordarse de ver a este niño correr. Así descrito, el recuerdo aparente de Jane no incluiría ninguna referencia a sí misma.

Aunque podríamos negar que los recuerdos aparentes de Jane tengan que parecer, en parte, versar sobre sí misma, no hay necesidad de hacerlo. Aunque sus recuerdos aparentes se presenten en el modo de la primera persona, Jane no tiene necesidad de asumir que, si no son ilusiones, tienen que ser recuerdos de sus *propias* experiencias. Aunque parezca acordarse de sí misma viendo el rayo hendido, podría concluir justificadamente que está cuasi-recordando una de las experiencias de Paul.

Algunos de los recuerdos aparentes de Jane no serían claramente de sus propias experiencias. Esto sería cierto en el caso de un recuerdo aparente de haberse afeitado *su* barba, mientras veía la cara de Paul en el espejo. En el caso de otros recuerdos aparentes, tal vez tendría que averiguar si fue ella o Paul la persona que tuvo una experiencia pasada. Y a veces esto sería imposible. Tal vez se viera obligada a decir, «Me parece acordarme con claridad de haber escuchado esa melodía. Pero no sé si fui yo o fue Paul el que la oyó». Cuando los recuerdos aparentes de Jane vienen a ella así, en un sentido son diferentes de los recuerdos aparentes del resto de nosotros. Como nosotros no tenemos cuasi-recuerdos de las experiencias pasadas de otras personas, nuestros recuerdos aparentes no nos vienen meramente en el modo de la primera persona. Nos vienen con la creencia de que, a no ser que sean ilusiones, tratan acerca de nuestras propias experiencias. Ahora bien, en el caso de los recuerdos de experiencias, esta es una creencia separable. Si, como Jane, tuviéramos cuasi-recuerdos de las experiencias pasadas de otras personas,

tales recuerdos aparentes dejarían automáticamente de estar combinados con esta creencia [15].

[15] Evans (2), p. 246, correctamente critica una mala descripción de la memoria en Parfit (1). Evans también argumenta contra «la posibilidad de una facultad que es, a la vez, igual que la memoria en dar a los sujetos conocimiento del pasado, y diferente de ella en que el contenido de los estados de memoria de ninguna manera llega hasta la cuestión de la propiedad del pasado al que se refieren» (p. 248). Esto entra en conflicto con mi afirmación de que los cuasi-recuerdos podrían proporcionar una manera de conocer las experiencias pasadas de otras personas. El argumento de Evans es este:

«Supongamos que quirúrgicamente "transferimos los recuerdos" del cerebro del sujeto S' al cerebro de un sujeto S, y supongamos que S no sabe que esto ha sucedido. Desde luego, S hará juicios sobre su pasado de la manera normal. Pero supongamos que descubre que él no fue F y que él no fue G,... —que en general no se puede confiar en su memoria como en un registro fiel del pasado—. Supongamos que, increíblemente, él entonces retrocede a hacer juicios en tiempo pasado *generales*: «Alguien fue F, y fue G...» *Estos juicios no podrían constituir conocimiento de ningún modo.* Hasta para resultar inteligible manifestándolos, S tendría que ofrecer lo que hubiera ocurrido realmente, o algo muy parecido, como hipótesis. Pero en absoluto podría decirse que sabía que era verdadero. Sería una pura conjetura. En consecuencia no se podría decir que sabe algo basado en ella. No hemos encontrado, por lo tanto, ninguna razón concluyente para abandonar la concepción de que nuestras ideas de nosotros mismos no permiten que se abra una brecha entre saber, en virtud de una operación de la memoria, que *alguien* vio un árbol ardiendo, y saber que fue *uno mismo* el que vio un árbol ardiendo» (pp. 244-5).

Acepto la afirmación de Evans de que su persona imaginaria S no tendría, en sus cuasi-recuerdos de las experiencias de S', conocimiento alguno de esas experiencias pasadas. Como escribe él mismo, «no es suficiente para constituir conocimiento el que una creencia verdadera sea causalmente dependiente de los hechos que la hacen verdadera». Estos cuasi-recuerdos no le dan a S conocimiento porque S no sabe nada de su causa. No se le ha dicho que los cirujanos han creado en su cerebro copias de las huellas mnémicas del cerebro de S'. Las cosas son diferentes para Jane. Ella conoce el modo en que sus cuasi-recuerdos de las experiencias de Paul son causalmente dependientes de esas experiencias. El argumento de Evans no demuestra que estos cuasi-recuerdos no le puedan dar a Jane conocimiento. No socava mi afirmación de que los cuasi-recuerdos podrían proporcionar un modo de conocer las experiencias pasadas de otras personas. Como sugiere Wiggins, podemos imaginar que esta clase de conocimiento es un fenómeno natural. Podría haber sido verdadero que los niños tuviesen cuasi-recuerdos de las experiencias de las vidas de sus dos padres, antes del momento de su concepción. Si tales cuasi-recuerdos adoptaran esta forma regular y restringida, podrían proporcionar conocimiento. [Wiggins (4), p. 145, tal vez atribuyéndole esta sugerencia a H. Ishiguro.]

Volvamos ahora a la objeción de Butler al criterio psicológico de identidad personal. Según esta objeción, no puede ser, ni siquiera en parte, la continuidad de memoria la que haga de una serie de experiencias la totalidad de las experiencias de una persona particular, desde el momento en que la memoria de la persona en cuestión presupone su identidad continua.

Según la interpretación que di arriba, la memoria presupone identidad porque, de acuerdo con nuestro concepto de memoria, sólo podemos recordar nuestras propias experiencias pasadas. Ahora podemos neutralizar esta objeción, pues podemos usar el concepto más amplio de cuasimemoria.

En nuestra exposición de nuestro criterio psicológico revisado, no deberíamos afirmar que, si tengo un cuasi-recuerdo fiel de una experiencia pasada, eso me convierte en la persona que tuvo esa experiencia. La vida mental de una persona puede incluir pocos cuasi-recuerdos de experiencias de la vida de otra persona, como en el caso imaginario de Jane y Paul. Nuestro criterio ignora esas pocas conexiones de cuasi-recuerdos. Y en vez de ello apelamos a cadenas parcialmente superpuestas de muchas de tales conexiones. Mi vida mental consiste en una serie de experiencias muy variadas. Entre ellas se incluyen incontables cuasi-recuerdos de experiencias anteriores. Las conexiones entre esos cuasi-recuerdos y esas experiencias anteriores se superponen parcialmente como las hebras de una cuerda. Hay *conexividad fuerte* de cuasimemoria si, a lo largo de cada día, el número de conexiones directas de cuasimemoria es como mínimo la mitad del número de conexiones que encontramos en la mayoría de las vidas reales. Hebras parcialmente superpuestas de conexividad fuerte nos proporcionan *continuidad de cuasimemoria.* Revisando a Locke, declaramos que la unidad de la vida de cada persona es en parte creada por esta continuidad. No estamos apelando ahora a un concepto que presuponga identidad personal. Como la continuidad de cuasimemoria no presupone identidad personal, puede ser parte de lo que la constituye. Puede ser parte de lo que me hace a mí ahora y a mí mismo en un momento diferente ser la misma persona. (Digo «parte» porque nuestro criterio también apela a las otras clases de continuidad psicológica.)

La objeción de Butler puede interpretarse de otra forma. Puede haber querido decir: «En la memoria somos directamente conscientes de nuestra propia identidad a través del tiempo, y también lo somos de que este es un hecho separado y adicional, que no puede consistir solamente en continuidad física y psicológica. Somos conscientes de que cada uno de nosotros es un sujeto de experiencias persistente, una entidad que existe separadamente y que no es nuestro cerebro ni nuestro cuerpo. Y somos conscientes de que nuestra propia existencia continua es, simplemente, la existencia continua de este sujeto de experiencias».

¿Esto es lo que nos dice nuestra memoria? ¿Somos directamente conscientes de la existencia de esta entidad separada, el sujeto de las experiencias? Los hay que han pensado que sí, y no sólo en la memoria, sino en todas nuestras experiencias.

81. El sujeto de experiencias

Reid escribe:

> «Mi identidad personal... implica la existencia continua de esa cosa indivisible que llamo mí mismo. Sea lo que sea este yo, es algo que piensa, delibera, resuelve, actúa y sufre. Yo no soy pensamiento, yo no soy acción, yo no soy sentimiento; soy algo que piensa, actúa y sufre» [16].

Hay un sentido en que esto es indudablemente cierto. Ni siquiera los reduccionistas niegan que existan las personas. Y, según el concepto que tenemos de ellas, las personas no son pensamientos ni actos. Son pensadoras y agentes. Yo no soy una serie de experiencias, sino la persona que *tiene* esas experiencias. Un reduccionista puede admitir que, en este sentido, una persona es *lo que tiene* experiencias, o el *sujeto de experiencias*. Esto es cierto por causa de nuestro modo de hablar. Lo que niega el reduccionista es que el sujeto de experiencias

[16] Reid, reeditado en Perry (1), p. 109.

sea una *entidad que exista separadamente*, distinta del cerebro y del cuerpo, y de una serie de sucesos físicos y mentales.

¿Es cierto que, en la memoria, somos directamente conscientes de lo que niega el reduccionista? ¿Es consciente cada uno de nosotros de que es un sujeto de experiencias que persiste en el tiempo, una entidad que existe separadamente, que no es su cerebro ni su cuerpo? ¿Es consciente cada uno de nosotros, por ejemplo, de que él es un Ego Cartesiano?

Sobre esta cuestión no se puede argumentar. Yo no creo que *yo* sea directamente consciente de que soy una entidad de esa clase. Y doy por sentado que no soy excepcional. Creo que nadie es directamente consciente de semejante hecho.

Supongamos que yo *fuera* consciente de ser tal entidad. Todavía quedaría una objeción contra la concepción cartesiana. Se ha dicho que yo no podría saber que esta entidad seguía existiendo. Como Locke y Kant plantearon [17], podría haber una serie de tales entidades que fueran psicológicamente continuas. Los recuerdos podrían pasarse de una a la siguiente como el testigo en una carrera de relevos. Y así con todos los demás elementos psicológicos. Dada la continuidad psicológica resultante, nosotros no seríamos conscientes de que una de estas entidades había sido sustituida por otra. Por consiguiente no podemos saber que tales entidades continúan existiendo.

Consideremos otra vez el caso de la línea secundaria, donde está claro que yo permanezco en la Tierra. Podría parecerle a una persona que acaba de tener estos dos pensamientos: «Cae la nieve. Así que tiene que hacer frío». Pero la verdad podría ser diferente. Esta persona es mi Réplica en Marte. Justo antes de que yo apretase el botón verde, pensé «Cae la nieve». Varios minutos después, mi Réplica se hace consciente de repente, en un cubículo parecido en Marte. Cuando se hace consciente, tiene recuerdos aparentes de vivir mi vida, y en especial parece recordar haber acabado de pensar,

[17] Véase Locke, Capítulo 27, Sección 13, reeditado en Perry (1), p. 101; y Kant, p. 342, especialmente la nota a pie de página a. Tomo el ejemplo de Wachsberg.

«Cae la nieve». Entonces piensa «Así que tiene que hacer frío». Mi Réplica en Marte estaría ahora en un estado mental exactamente como el mío cuando acababa de tener estos pensamientos. Cuando mi Réplica se halla en este estado mental, creería que estos dos pensamientos fueron tenidos por el mismo ser pensante, ella misma. Pero esto sería falso. Yo tuve el primer pensamiento, y mi Réplica nada más que tuvo el segundo.

Este ejemplo es imaginario. Pero parece demostrar que no podríamos decir, partiendo del contenido de nuestras experiencias, si somos realmente conscientes de la existencia continua de un sujeto de experiencias que existe separadamente. Como mucho, lo que tenemos son estados mentales como los de mi Réplica. Mi Réplica piensa falsamente que acaba de tener dos pensamientos. No es consciente de la existencia continua de una entidad que existe separadamente: el pensador de esos pensamientos. Es consciente de algo menos, la continuidad psicológica entre su vida y la mía. De la misma manera, cuando hemos tenido una serie de pensamientos, de lo que podemos ser conscientes, como mucho, es de la continuidad psicológica de nuestra corriente de conciencia. Los hay que afirman que somos conscientes de la existencia continua de sujetos de experiencias existiendo separadamente. Como Locke y Kant plantearon, y como nuestro ejemplo parece demostrar, esa conciencia no puede de hecho distinguirse de nuestra conciencia de la mera continuidad psicológica. Nuestras experiencias no nos dan razón alguna para creer en la existencia de esas entidades. A no ser que tengamos otras razones para creer en ella, deberíamos rechazar esta creencia.

Esta conclusión no es, como opinan algunos, burdamente verificacionista. No estoy dando por supuesto que sólo podría ser verdadero en definitiva aquello que pudiéramos conocer. Mis observaciones parten de una asunción diferente. Estoy discutiendo una afirmación general sobre la existencia de una clase concreta de cosa. Se dice que se trata de una entidad que existe separadamente, que es distinta de nuestros cerebros y nuestros cuerpos. Yo digo que, si no contamos con razones para creer que tales entidades existen, deberíamos rechazar esta creencia. No digo, como los verificacionistas, que esta creencia sea un sinsentido. Mi afirmación es simplemente como la de que, puesto que no tenemos razones para creer en la existencia de las ninfas del agua o de los unicornios, deberíamos rechazar estas creencias [18].

Aunque no seamos directamente conscientes de la existencia de estas entidades, hay quienes afirman que podemos deducir su existencia de cualquiera de nuestras experiencias. El más famoso de los que hizo esta declaración fue Descartes. Cuando preguntó si había algo que no se pudiera dudar, su respuesta fue que no podía poner en duda su propia existencia. Esta quedaba de manifiesto en el mismo acto de dudar. Y, además de dar por supuesto que todo pensamiento tiene que tener un pensador, Descartes dio por supuesto que un pensador tiene que ser un ego puro, una sustancia espiritual. Un Ego Puro Cartesiano es el caso más claro de entidad que existe separadamente, distinta del cerebro y del cuerpo [19].

Lichtenberg afirmó que Descartes se equivocó justo en lo que pensaba que era lo más cierto. No debería haber afirmado que un pensador tiene que ser una entidad que existe separadamente. Su famoso *Cogito* no justificaba esta creencia. No debería haber afirmado, «Pienso, luego existo». Aunque esto es verdadero, resulta engañoso. Descartes podría haber dicho en vez de eso, «Se piensa: el pensamiento tiene lugar». O también, «Esto es un pensamiento, por tanto, como mínimo, se piensa un pensamiento» [20].

[18] Los No Reduccionistas tienen otros varios argumentos que, en una discusión más larga, yo necesitaría tratar de responder. Además de Swinburne, véase en particular Lewis (4), (5), y (6), y Madell. Espero discutir estos argumentos en otro lugar.

[19] Descartes, p. 101: «Me di cuenta de que, mientras que de este modo yo deseaba pensar que todas las cosas eran falsas, *era absolutamente esencial que el "yo" que pensaba esto debiese ser un algo*, y observando que esta verdad, "pienso, luego existo", era tan cierta y tan segura que todas las más extravagantes suposiciones presentadas... eran incapaces de hacerla tambalear, llegué a la conclusión de que podía admitirla sin escrúpulo como el primer principio de la Filosofía que estaba buscando». (El primer énfasis mío.)

[20] Lichtenberg, p. 412. (En el original: «*Es denkt*, sollte man sagen, so wie man sagt: *es blitzt*» [«Se debiera decir *piensa*, como se dice: *relampaguea*»].) Para deletrear el aforismo: de la verdad de «pienso luego existo», Descartes *no* debería haber asumido que «era absolutamente esencial que el "yo" que pensaba esto debiera ser un algo».

Puesto que adscribimos pensamientos a pensadores, podemos afirmar verdaderamente que los pensadores existen. Pero no podemos deducir, del contenido de nuestras experiencias, que un pensador sea una entidad que existe separadamente. Y, como sugiere Lichtenberg, puesto que no somos entidades que existen separadamente, podríamos describir exhaustivamente nuestros pensamientos sin afirmar que tienen pensadores. Podríamos describir exhaustivamente nuestras experiencias, y las conexiones entre ellas, sin afirmar que son tenidas por un sujeto de experiencias. Podríamos dar lo que llamo una descripción *impersonal*.

Como he dicho, hay autores que rechazan *tanto* esta última afirmación reduccionista *cuanto* la concepción cartesiana. No creen en Egos Puros Cartesianos. Y no piensan que una persona sea ninguna otra clase de entidad que exista separadamente. Piensan que la existencia de una persona consiste sencillamente en la existencia de su cerebro y de su cuerpo, y en el llevar a cabo sus actos, y en la ocurrencia de otros diversos sucesos, físicos y mentales. Pero estos autores dicen que no podemos referirnos a experiencias concretas, ni describir las conexiones que se dan entre ellas, como no nos refiramos a la persona que las tiene. Según su modo de ver las cosas, la unidad de una vida mental no puede explicarse de manera impersonal.

Strawson discute un argumento a favor de esta concepción, un argumento sugerido por Kant. Este argumento afirma que no podríamos tener conocimiento del mundo que nos rodea como no pensáramos que nosotros mismos somos personas, con una conciencia de nuestra identidad a lo largo del tiempo. Shoemaker plantea un argumento similar. Si estos argumentos son correctos, podrían refutar mi afirmación de que podríamos redescribir nuestras vidas de un modo impersonal. Como estos argumentos se sitúan a un nivel muy abstracto, tengo la esperanza de discutirlos en otra parte [21ª].

Williams discute una objeción más simple a la descripción impersonal [21]. Esta objeción está dirigida a Lichtenberg. Como señala Williams, el sustituto sugerido por Lichtenberg para el *Cogito*

[21ª] Véase Strawson (2) y Shoemaker (2).
[21] Williams (4), pp. 95-100.

de Descartes no tiene por qué ser completamente impersonal. No tiene por qué ser, «Se piensa: el pensamiento tiene lugar». Sino que podría ser, «Se piensa: estoy pensando». Como el sujeto de experiencias es mencionado aquí sólo en el *contenido* del pensamiento, esta frase no adscribe este pensamiento a un pensador.

Williams señala entonces que, si varios pensamientos fueran expresados de esta manera, haría falta dejar claro si los mismos ocurrían dentro de la misma vida o en vidas diferentes. Porque esto no estaría claro si todos estos pensamientos empezaran con el giro «Se piensa: ...». Él considera «(T10) Se piensa en el lugar A: ...», pero rechaza este giro. Continúa diciendo: «... se necesita un sustituto menos figurativo para "en el lugar A" en la formulación de la ocurrencia del pensamiento —y es natural concluir que nada que no sea un nombre personal, o algo por el estilo, servirá como sustituto, de forma que T10 habrá de ser reemplazada por

(T11) A piensa: ...

En este punto... el programa de introducir formulaciones impersonales habrá finalmente fracasado».

Williams sugiere la respuesta a esta objeción. Como escribe, «Podría haber posiblemente algún sustituto para los "lugares" figurativos que sirviera a los propósitos de una relativización efectiva, pero que no fuera tan lejos como para introducir un sujeto que piensa». Hay muchos sustitutos así. Dos podrían ser estos:

> En la vida concreta que contiene el pensar del pensamiento que se expresa por la proferencia de esta sentencia, se piensa:...

o

> En la vida concreta que ahora es causalmente dependiente del cuerpo A, se piensa:...

Lichtenberg necesitaría entonces explicar la unidad de la vida de una persona de un modo impersonal. Primero podría ajustar nuestro concepto de cuasimemoria. Podría afirmar que un recuerdo aparente es un cuasi-recuerdo fiel si

(1) el recuerdo aparente es de una experiencia pasada determinada,
(2) esta experiencia ocurrió,

y

(3) el recuerdo aparente es causalmente dependiente, del modo correcto, de esta experiencia.

Tendría que demostrar además que la clase correcta de causa puede describirse de un modo que no presuponga identidad personal. Podría entonces apelar a las otras clases de continuidad psicológica, tales como la que se da entre la formación de una intención y el acto posterior en la que la intención se lleva a cabo. Todavía tengo que demostrar que estas otras continuidades, y sus causas, se pueden describir de formas que no presuponen identidad personal. Puesto que se pueden describir así, como demuestro en la Sección 89, podrían también describirse de un modo impersonal. Hay que mencionar a las personas al describir el *contenido* de pensamientos, deseos y otras innumerables experiencias. Pero, como señala Williams, tales descripciones no afirman que estas experiencias son *tenidas* por personas. Y, sin hacer esta afirmación, podríamos describir las interrelaciones entre todos los sucesos mentales y físicos que, juntos, constituyen la vida de una persona concreta.

De forma que la objeción de Lichtenberg a Descartes tiene vigencia. Podemos referirnos a diferentes pensamientos, y describirlos, y además describir las relaciones que se dan entre los mismos, sin adscribirlos a sujetos pensantes. En efecto, adscribimos los pensamientos a sujetos pensantes. Porque hablamos del modo en que lo hacemos, Descartes pudo verdaderamente afirmar, «Pienso luego existo». Pero Descartes no demostró que un pensador tenga que ser una entidad que existe separadamente, distinta de un cerebro y de un cuerpo, y de varios sucesos mentales y físicos [22].

[22] Véanse, por ejemplo, las objeciones de Williams en Williams (4).

82. Cómo podría haber sido verdadera una Concepción No Reduccionista

Hay autores que sostienen que el concepto de Ego Cartesiano es ininteligible. Yo cuestiono esta opinión. Pienso que podría haber habido evidencia en apoyo de la concepción cartesiana.

Por ejemplo, podría haber habido evidencia en apoyo de la creencia en la reencarnación. Evidencia como esta: una mujer japonesa podría afirmar que se acuerda de haber vivido una vida de cazadora y guerrera celta en la Edad de Bronce. Basándose en sus recuerdos aparentes ella sería capaz de hacer muchas predicciones que podrían ser comprobadas por los arqueólogos. Podría por ejemplo decir que se acuerda de haber tenido una pulsera de bronce, con la forma de dos dragones luchando. Y podría decir que recuerda haberla enterrado junto a un megalito determinado, justo antes de la batalla en que la mataron. Los arqueólogos podrían encontrar ahora una pulsera como esa enterrada en el lugar indicado, y sus instrumentos podrían demostrar que aquí la tierra no ha sido removida durante al menos 2000 años. La mujer japonesa podría hacer muchas otras predicciones semejantes, y todas ellas ser verificadas.

Supongamos además que hay otros muchísimos casos en los que gente que está viva hoy afirma recordar haber vivido vidas pasadas muy concretas, y se nos proporciona predicciones similares que todas se verifican. Y esto llega a pasar con la mayoría de la gente en la población mundial. Si hubiera evidencia suficiente de este tipo, y no hubiese otro modo posible de explicar cómo la mayoría de nosotros podía conocer tantos hechos detallados del pasado lejano, podríamos tener que admitir que tenemos cuasi-recuerdos fieles de esas vidas pasadas. Podríamos tener que concluir que la mujer japonesa tiene una forma de conocer la vida de una guerrera celta de la Edad de Bronce que es como su memoria de su propia vida.

Podría descubrirse además que no hay continuidad física entre la guerrera celta y la mujer japonesa. Tendríamos por consiguiente que abandonar la creencia de que el portador de la memoria es el cerebro. Tendríamos que asumir que la causa de estos cuasi-recuerdos es algo puramente mental. Tendríamos que asumir que hay una

entidad puramente mental, que estaba implicada de algún modo en la vida de la guerrera celta, y que ahora está de algún modo implicada en la vida de la mujer japonesa, y que ha seguido existiendo durante los miles de años que separan las vidas de estas dos personas. Un Ego Cartesiano es precisamente una entidad semejante. Si hubiera suficiente evidencia de la reencarnación, tendríamos una razón para pensar que realmente hay tales entidades. Y entonces podríamos concluir razonablemente afirmando que tal entidad es lo que cada uno de nosotros realmente es.

Esta clase de evidencia no daría apoyo directo a la afirmación de que los Egos Cartesianos tienen las otras propiedades especiales que los cartesianos creen que tienen. Por ejemplo, no demostraría que la existencia continua de estos egos es todo-o-nada. Pero podría haber habido evidencia en apoyo de esta tesis. Podrían haberse dado varias clases o grados de daño en el cerebro de una persona que no alteraran a la persona de ningún modo fundamental, mientras que otras clases o grados de daño parecieran producir una persona completamente nueva, de ninguna manera psicológicamente continua con la persona original. Algo parecido podría suceder con las diversas clases de enfermedad mental. Podríamos haber llegado a la conclusión, en general, de que estos tipos de interferencia o bien no hicieron nada en absoluto para destruir la continuidad psicológica, o bien la destruyeron por completo. Podría haberse demostrado imposible encontrar o producir casos intermedios, en los cuales la conexividad psicológica se diera en grados reducidos.

¿Contamos con buenas evidencias para la creencia en la reencarnación? ¿Y contamos con evidencia para creer que la continuidad psicológica depende principalmente, no de la continuidad del cerebro, sino de la continuidad de alguna entidad diferente, que o bien existe no mermada o bien no existe en absoluto? En efecto, no contamos con este tipo de evidencia descrito arriba. Aunque podamos entender el concepto de un Ego Puro Cartesiano o de una sustancia espiritual, no disponemos de evidencia que nos haga creer que tal entidad existe. Ni tampoco que nos haga creer que una persona es alguna otra clase de entidad que existe separadamente. Y en cambio tenemos mucha evidencia no sólo para pensar que el portador

de la continuidad psicológica es el cerebro, sino también para creer que la conexividad psicológica podría darse en cualquier grado reducido [23].

He concedido que la versión mejor conocida de la Concepción No-Reduccionista, la que sostiene que somos Egos Cartesianos, puede tener sentido. Y he sugerido que, si los hechos hubiesen sido muy diferentes, se habría dado una evidencia suficiente como para creer en esta concepción. Los hay que creen en Egos Cartesianos y no los conectan, de estas maneras, con hechos observables, sino que aceptan la posibilidad descrita por Locke y Kant. Según su opinión, el Ego Cartesiano que soy podría dejar de existir repentinamente y ser reemplazado por otro Ego. Este nuevo Ego podría «heredar» todas mis características psicológicas, como en una carrera de relevos. Según esta *Concepción Cartesiana Monótona*, mientras tú estás leyendo esta página podrías dejar de existir de repente y tu cuerpo ser ocupado por una nueva persona que es justo como tú. Si así ocurriera, nadie notaría ninguna diferencia. Nunca habría evidencia alguna, ni pública ni privada, que mostrara si esto ocurre o no, y, en caso afirmativo, con cuánta frecuencia. Por eso no podemos ni siquiera afirmar que no es probable que ocurra. Y hay otras posibilidades. Según esta concepción, la historia podría haber ido justo como ha ido, excepto que yo fui Napoleón y Napoleón fue yo. Lo que no equivale a decir que Derek Parfit podría haber sido Napoleón. La idea es más bien que yo soy un Ego Cartesiano, Napoleón otro, y estos dos Egos podrían haber «ocupado» cada uno el lugar del otro [24].

Cuando se priva de este modo a la creencia en Egos Cartesianos de cualquier conexión con hechos públicamente observables o privadamente introspeccionables, la acusación de que es ininteligible se

[23] Es demasiado dogmático decir que no tenemos *ninguna* evidencia a favor de una Concepción No Reduccionista. La evidencia que tenemos, observó una vez C. D. Broad con buen juicio, apoya como mucho la siguiente conclusión: si una persona aparentemente sana y seria afirmara tener *mejor* evidencia de la misma clase, su afirmación no debería ser simplemente ignorada.

[24] Estas observaciones derivan de «Imagination and the Self» [«La imaginación y el yo»], de Williams, reeditado en Williams (2).

hace más plausible. Y no está claro que los cartesianos puedan evitar esta versión de su tesis. No está claro que puedan negar la posibilidad descrita por Locke y Kant. Pero basta con repetir que tenemos razones suficientes para rechazar esta tesis.

83. El argumento de Williams contra el Criterio Psicológico

He defendido el criterio psicológico de dos maneras. He afirmado, y en parte mostrado, que podemos describir la continuidad psicológica de modo que no presuponga identidad personal. Y he afirmado que, juzgando a partir de la evidencia con la que contamos, el portador de esta continuidad no es una entidad que exista separadamente del cerebro y el cuerpo de la persona.

A continuación voy a considerar otra objeción al criterio psicológico. Es la planteada por Williams [25], una objeción que parece demostrar que si el cerebro de una persona sigue existiendo, y sirviendo de soporte a su conciencia, la persona seguirá existiendo, por grandes que sean las rupturas en la continuidad psicológica de su vida mental.

Aquí tenemos una versión más simple de esta objeción. Consideremos

> *El Ejemplo de Williams.* Soy prisionero de un neurocirujano cruel, que va a intentar interrumpir mi continuidad psicológica manipulando mi cerebro. Estaré consciente mientras me opera, y me dolerá mucho. Por eso me horroriza pensar en lo que me espera.

El cirujano me dice que, mientras yo esté sufriendo, él hará varias cosas. Primero activará unos electrodos que me provocarán amnesia. Perderé de golpe todos los recuerdos de mi vida hasta el comienzo mismo de mi dolor. ¿Acaso me da esto menos razones para temer lo que va a hacerme? ¿Puedo figurarme que cuando el ciruja-

[25] En Williams (8).

no le dé a este interruptor el dolor que siento cesará de repente? Casi seguro que no. El dolor podría llegar a ocupar mi mente hasta tal extremo que yo ni notara la pérdida de todos esos recuerdos.

Luego el cirujano me dice que, mientras yo esté todavía sufriendo, el girará otro interruptor que hará que me crea Napoleón, y me proporcionará recuerdos aparentes de la vida de Napoleón. ¿Puedo figurarme que esto hará que mi dolor desaparezca? La respuesta natural es otra vez No. Para dar apoyo a esta respuesta podemos de nuevo suponer que mi dolor no me dejará notar nada. No notaré el ponerme a creer de repente que soy Napoleón, no notaré que adquiero todo un nuevo conjunto de recuerdos aparentes. Cuando el cirujano gire el segundo interruptor, no habrá en absoluto ningún cambio en nada de lo que soy consciente. Los cambios serán puramente disposicionales. Sólo llegaría a ocurrir que, si el dolor desapareciera hasta el punto de que yo fuese capaz de pensar, contestaría a la pregunta «¿Quién eres tú?» con el nombre «Napoleón». De manera similar, si el dolor desapareciera, yo comenzaría entonces a tener recuerdos aparentes ilusorios, como por ejemplo los de haber pasado revista a la Guardia Imperial, o haber llorado de frustración por la catástrofe de 1812. Si el darle al segundo interruptor sólo conllevase tales cambios en mis disposiciones, yo no tendría ninguna razón para esperar que esto iba a hacer que mi dolor cesara.

Luego el cirujano me dice que, durante mi suplicio, presionará a continuación un tercer interruptor, y que eso cambiará mi carácter de forma que se convierta en el mismísimo carácter de Napoleón. Una vez más, no parece que tenga yo ninguna razón para esperar que el movimiento de este interruptor ponga fin a mi dolor. Como mucho, podría traerme algún alivio, en el caso de que el carácter de Napoleón, comparado con el mío, fuese un carácter con más fortaleza.

En este caso imaginario, nada de lo que se me dice parece darme razones para esperar que, durante el suplicio al que me van a someter, yo dejaré de existir. Y parezco tener las mismas razones que antes para temer todo este tormento. No parece que estas razones sean eliminadas por las otras cosas que tengo que temer —perder mis recuerdos, volverme loco, volverme como Napoleón y creer que

soy él—. Como Williams afirma, este argumento parece demostrar que puedo tener razones para temer un dolor futuro, sean los que sean los cambios psicológicos que precedan a este dolor. Incluso tras todos estos cambios, seré yo el que sienta el dolor. Si es así, el criterio psicológico de identidad personal está equivocado. En este caso imaginario, entre yo ahora y yo mismo después del suplicio, no habría continuidad de memoria, ni de carácter ni de nada por el estilo. Lo que conlleva mi continuar existiendo, por consiguiente, no puede ser semejante continuidad [26].

Puede objetarse que, si permanezco consciente durante todo el suplicio, por lo menos habrá un tipo de continuidad psicológica. Aunque pierda todos mis recuerdos de mi vida pasada, yo tendría recuerdos de mis tremendos dolores. Sobre todo, seguiría teniendo recuerdos a corto plazo de los últimos momentos más recientes, de eso que a veces se llama *presente especioso*. Durante todo mi suplicio habría una cadena parcialmente superpuesta de tales recuerdos.

Para anular esta objeción podemos añadir una nueva característica al caso. Después de haber perdido todos mis otros recuerdos, por un momento me dejan inconsciente. Cuando recobro la conciencia, *no* tengo recuerdos. A medida que prosigue el suplicio, yo tendría nuevos recuerdos. Pero no habría continuidad de memoria más allá de mi momento de inconsciencia.

Puede objetarse ahora que he descrito esta historia en unos términos que dan por establecido lo que se pretende demostrar. Y es que sugerí que, cuando se me hace perder mis recuerdos, yo, por causa del dolor, sería incapaz de notar ningún cambio. Esta descripción asume que, tras la pérdida de mis recuerdos, la persona que sufre el tormento todavía sería yo. Tal vez sea cierto que, en este punto, yo dejaría de existir, y una nueva persona comenzaría a existir en mi cuerpo.

Williams contestaría que, aunque mi descripción asuma que yo seguiría existiendo, se trata de la asunción que resulta abrumadoramente convincente. Es el defensor del criterio psicológico el que

[26] Habría ciertas clases de continuidad *no distintiva*, tales como la memoria continua de cómo caminar y correr. El Criterio Psicológico no debería apelar a estas clases de continuidad psicológica.

tiene que demostrar que no está justificada. Y esto sería difícil. Es difícil de creer que, si me hacen perder mis recuerdos mientras estoy pasando dolores terribles, esto va a hacerme dejar de existir en mitad del tormento. Y es difícil de creer que el cambio de mi carácter tendría este efecto.

El argumento de Williams parece refutar el criterio psicológico. Parece demostrar que la concepción verdadera es la del criterio físico. Según ella, si el cerebro y el cuerpo siguen existiendo, y sirviendo de soporte a la conciencia, la persona seguirá existiendo, por muy grandes que sean las rupturas en la continuidad psicológica de su vida mental.

84. El Espectro Psicológico

Voy a revisar ahora el argumento de Williams. Más tarde se verá por qué vale la pena hacerlo.

Williams discute un caso particular en que, tras unos pocos cambios, deja de haber continuidad psicológica. Yo discutiré un *espectro*, o gama de casos, de los que cada uno es muy similar a sus vecinos. Estos casos suponen todos los grados posibles de conexividad psicológica. Lo llamo el *Espectro Psicológico*.

En el caso situado en el extremo lejano, el cirujano apretaría simultáneamente muchísimos interruptores. Y esto provocaría que no hubiese conexiones psicológicas entre yo mismo y la persona resultante. Esta persona sería completamente como Napoleón.

En los casos situados en el extremo cercano, el cirujano sólo daría a unos cuantos interruptores. Si apretara sólo el primer interruptor, simplemente me haría perder unos pocos recuerdos y adquirir unos pocos recuerdos aparentes que encajan en la vida de Napoleón. Si le diera a los dos primeros interruptores, yo simplemente perdería unos pocos recuerdos más, adquiriendo unos pocos más de los nuevos recuerdos aparentes. Sólo si apretara todos los interruptores perdería yo la totalidad de mis recuerdos, y adquiriría un conjunto completo de ilusiones napoleónicas.

Algo parecido ocurre con los cambios de mi carácter. Cualquier interruptor por sí solo no causaría nada más que un pequeño cam-

bio. Así que, si voy a ser como Napoleón, tengo que ponerme de peor humor y me tiene que dejar de impresionar el ver cómo matan a la gente. Estos serían los únicos cambios que se producirían si se apretaran los dos primeros interruptores.

En esta versión revisada del argumento, que incluye muchísimos casos diferentes, tenemos que decidir cuáles son los casos en los que yo sobreviviría. En el caso del extremo cercano, el cirujano no hace nada. En el segundo caso, yo simplemente perdería unos pocos recuerdos, tendría unas pocas ilusiones y me pondría de mal humor. Está claro que, en este caso, yo sobreviviría. En el tercer caso los cambios sería sólo ligeramente mayores. Y esto sucede con cualesquiera de dos casos vecinos en la gama. Es difícil de creer que yo sobreviviera en uno de estos casos, pero que, en el siguiente, yo dejara de existir. Que siga existiendo no podemos pensar convincentemente que dependa de si pierdo sólo unos cuantos recuerdos más, y tengo unos pocos recuerdos ilusorios más, y de si mi carácter ha cambiado de algún modo mínimo. Si ninguno de estos cambios pequeños podría hacer que dejase de existir, yo seguiría existiendo en todos esos casos. Seguiría existiendo incluso en el caso situado en el extremo lejano del espectro. Pero, en ese caso, entre mí mismo ahora y la persona resultante *no* habría conexiones psicológicas.

Puede objetarse:

> En esta forma revisada, el argumento se parece sospechosamente a los que están implicados en el *Problema Sorites*, o la *Paradoja del Montón*. En ellos somos conducidos, a través de lo que parecen pasos inocentes, a conclusiones absurdas. Tal vez aquí ocurra lo mismo.
>
> Supongamos que afirmamos que la eliminación de un único grano no puede transformar un montón de arena en algo que no es un montón. Alguien empieza con un montón de arena, del que va quitando grano a grano. Nuestra afirmación anterior nos fuerza a admitir que, tras cada cambio, todavía tenemos un montón, aunque el número de granos llegue a tres, a dos o a uno. Pero sabemos que hemos ido a parar a una conclusión falsa. Un grano no es un montón.
>
> En tu apelación al espectro psicológico, dices que ningún pequeño cambio podría hacerte dejar de existir. Haciendo pequeños cambios en número suficiente, el cirujano podría provocar que la persona resultante no estuviera de ningún modo psicológicamente conectada contigo. El argumento te forzó a concluir que la persona resultante serías tú. Pero esta conclusión puede que sea igual de falsa que la conclusión sobre el grano de arena.

Para defender esta versión del argumento de Williams, no necesito resolver el Problema Sorites. Bastará con hacer los comentarios siguientes.

Cuando consideramos montones, todos pensamos que hay casos fronterizos. ¿Son dos granos de arena un montón? ¿Lo son cuatro, ocho, dieciséis? Tal vez no sepamos contestar a todas estas preguntas. Pero no pensamos que esto sea resultado de nuestra ignorancia. No pensamos que todas estas preguntas tengan que tener una respuesta. Sabemos que el concepto de montón es vago, con fronteras vagas. Y cuando el Argumento Sorites se aplica a montones, nos quedamos contentos resolviendo el problema con una *estipulación*: una decisión arbitraria sobre cómo usar la palabra «montón». Podríamos decidir que no diremos que nueve granos forman un montón, pero que llamaremos montón a cualquier colección de diez o más granos. Entonces habremos abandonado una de las premisas del argumento. Según nuestro nuevo y más preciso concepto, la eliminación de un grano único puede transformar un montón de arena en algo que no es un montón. Esto ocurre con la eliminación del grano décimo último.

Cuando se aplica a otros temas, como por ejemplo el color fenoménico, el Argumento Sorites no puede ser descartado con tanta facilidad [27]. Ni tampoco se lo puede descartar de una manera que parezca plausible cuando el argumento se aplica a la identidad per-

[27] Como antes, para una discusión de estos argumentos véase Dummett, Peacocke (1), Forbes (2), y Sainsbury. Véase también C. Wright, «On the Coherence of Vague Predicates» [«Sobre la coherencia de los predicados vagos»], *Synthese*, 1975. Para una discusión más extensa del argumento en tanto que aplicado a las personas, véanse especialmente Unger (1), Unger (2), y los demás trabajos de P. Unger en esta excelente serie. Todavía no he tenido tiempo de considerar si la solución sugerida por Peacocke, Forbes y Sainsbury, la que apela a *grados de verdad*, contrarresta los argumentos de Unger.

sonal. La mayoría de nosotros piensa que nuestra propia existencia continua no es, en varios puntos de importancia, como la existencia continua de un montón de arena.

Volvamos a considerar la gama de casos en el Espectro Psicológico. Como el ejemplo de Williams, estos casos proporcionan un argumento contra el Criterio Psicológico. Este criterio es una versión de la Concepción Reduccionista. Y un reduccionista podría decir:

> El argumento asume que, en cada uno de estos casos, la persona resultante sería o no sería yo. Pero esto no es así. La persona resultante sería yo en unos pocos primeros casos. En el último caso no sería yo. En muchos de los casos intermedios, ninguna de las dos respuestas sería correcta. Siempre puedo preguntar, «¿Estoy a punto de morir?». «¿Habrá alguna persona viva que será yo?». Pero, en los casos situados en la mitad del espectro, no hay respuesta a esta pregunta.
>
> Aunque no haya respuesta a esta pregunta, yo podría saber exactamente lo que ocurrirá. La pregunta es aquí *vacía*. En cada uno de estos casos yo podría saber hasta qué grado estaría yo psicológicamente conectado con la persona resultante. Y yo podría saber qué conexiones particulares se darían o no se darían. Si conociera estos hechos, lo sabría todo. Aún puedo preguntar si la persona resultante sería *yo*, o simplemente sería *alguien distinto* que es en parte como yo. En ciertos casos, estas son dos posibilidades diferentes, una de las cuales tiene que ser verdadera. Pero en *estos* casos no se trata de dos posibilidades diferentes. Son simplemente dos descripciones del mismo curso de sucesos.

Estos comentarios son análogos a los que aceptamos cuando nos referimos a montones. No pensamos que cualquier colección de arena tenga que ser o no ser un montón. Sabemos que hay casos límite, en los que no hay una respuesta obvia a la pregunta «¿Hay todavía un montón?». Pero no pensamos que, en estos casos, tenga que *haber* una respuesta, que tenga que ser Sí o No. Pensamos que en esos casos se trata de una pregunta vacía. Aunque no la respondamos, lo sabemos todo.

Como Williams dice, cuando se aplican a nuestra propia existencia, estas observaciones parecen increíbles. Supongamos que estoy a punto de sufrir una operación en la mitad de este espectro. Sé que la persona resultante sufrirá. Si no sé si yo seré o no seré la persona que sufrirá, y ni siquiera sé si todavía estaré vivo, ¿cómo puedo pensar que *sé* exactamente lo que va a ocurrir? No conozco la respuesta a las preguntas más importantes. Es muy difícil pensar que se trata de preguntas vacías.

La mayoría de nosotros piensa que no somos como los montones, porque nuestra identidad tiene que ser determinada. Pensamos que, incluso en los «casos límite», la pregunta «¿Estoy a punto de morir?» tiene que tener una respuesta. Y, como mantiene Williams, creemos que la respuesta tiene que ser la más simple, Sí o No. Si alguien va a vivir, y a sufrir terribles dolores, o será o no será yo. Una de las dos cosas tiene que ser verdadera. Y no podemos darle ningún sentido a una tercera alternativa, como la de que la persona sufriente será *en parte* yo. Me puedo imaginar que sufro sólo parcialmente, en la medida en que pierdo y recobro la conciencia. Pero si alguien va a ser completamente consciente del dolor, esta persona no puede ser yo parcialmente.

La Concepción Reduccionista proporcionaría una respuesta al argumento de Williams. Cuando Williams da su versión de este argumento, rechaza esta concepción. En vez de ello concluye que, si sigue existiendo mi cerebro, y sigue siendo el cerebro de una persona viva, yo seré esa persona. Esto sería así aunque, entre yo mismo ahora y yo mismo después *no* hubiera conexiones psicológicas. Después de exponer su argumento, Williams dice que esta conclusión «quizás» sea errónea, «pero que hace falta que se nos demuestre lo que es erróneo en ella» [28].

85. El Espectro Físico

Una objeción es la de que un argumento similar se aplica a la continuidad física. Consideremos otra gama de casos posibles: el *Espectro Físico*. Estos casos involucran todos los diferentes grados posibles de continuidad física.

[28] Williams (2), p. 63.

En el caso situado en el extremo cercano de este espectro, habría después una persona que sería plenamente continua conmigo como yo soy ahora, tanto física como psicológicamente. En el caso situado en el extremo lejano, habría después una persona que sería psicológica pero no físicamente continua conmigo como yo soy ahora. El extremo lejano es como el caso del teletransporte. El extremo cercano es el caso normal de la existencia continua.

En un caso próximo al extremo cercano, los científicos reemplazarían el 1% de las células de mi cerebro y de mi cuerpo con duplicados exactos. En el caso de la mitad del espectro, reemplazarían el 50%. En un caso próximo al extremo lejano, reemplazarían el 99%, dejando solamente el 1% de mi cerebro y mi cuerpo originales. En el extremo lejano, el «reemplazo» llevaría consigo la destrucción completa de mi cerebro y de mi cuerpo, y la creación a partir de nueva materia orgánica de una Réplica mía.

Lo que es importante en este último caso no es sólo que el cerebro y el cuerpo de mi Réplica estarían compuestos enteramente de materia nueva. Como expliqué antes, esto podría ocurrir de un modo que no destruyese mi cerebro y mi cuerpo. Podría ocurrir así en el caso de que se diera una larga serie de pequeños cambios en la materia de mi cuerpo, durante los que mi cerebro y mi cuerpo siguieran existiendo y funcionando normalmente. Esto sería como el barco que se transforma en uno compuesto de nuevos trozos de madera tras cincuenta años de reparaciones pedazo a pedazo. En ambos casos, el cambio completo de la identidad de los componentes no interrumpe la continuidad física. Las cosas son diferentes en el caso situado en el extremo lejano del espectro físico. Aquí no hay continuidad física, puesto que mi cerebro y mi cuerpo son completamente destruidos, y sólo después los científicos crean a mi Réplica, a partir de materia nueva.

Los primeros pocos casos en esta gama se cree ahora que son técnicamente posibles. Se han trasplantado con éxito porciones de tejido cerebral del cerebro de un mamífero al de otro. Y lo que se trasplanta podría ser una parte del cerebro que, en todos los individuos, fuera suficientemente similar. Esto podría permitir a los cirujanos proporcionar sustitutos que funcionen de algunas partes

lesionadas del cerebro. Estos trasplantes reales se vio que eran más fáciles que los más familiares trasplantes de riñón o de corazón, desde el momento en que un cerebro no parece «rechazar» el tejido trasplantado de la misma manera en que el cuerpo rechaza los órganos trasplantados [29]. Aunque los primeros pocos casos en esta gama son posibles incluso ahora, la mayoría de los casos seguirá siendo imposible. Pero tal imposibilidad será una imposibilidad meramente técnica. Como yo uso estos casos sólo para descubrir lo que creemos, esta imposibilidad no importa.

Supongamos que pensamos que, en el extremo lejano de este espectro, mi Réplica no sería yo. Simplemente sería alguien diferente que es exactamente como yo. En el extremo cercano de este espectro, donde no habría reemplazo, la persona resultante sería yo. ¿Qué debería esperar si lo que va a suceder es un caso intermedio? Si reemplazaran sólo el 1%, ¿dejaría yo de existir? Esto no es plausible, porque yo no necesito la totalidad de mi cerebro y de mi cuerpo. Pero, ¿qué ocurre con los casos en que reemplazan el 10%, el 30%, el 60%, o el 90%?

Esta gama de casos cuestiona el criterio físico, que es otra versión de la concepción reduccionista. Imagínate que estás a punto de sufrir una de estas operaciones. Podrías tratar de creer en esta versión del reduccionismo. Podrías decirte a ti mismo:

> En todo caso de esta gama que ocupe una posición central, la pregunta «¿Estoy a punto de morir?» carece de respuesta. Pero sé lo que va a ocurrir. Un determinado porcentaje de mi cerebro y de mi cuerpo será reemplazado con duplicados exactos de las células existentes. La persona resultante será psicológicamente continua conmigo como soy ahora. Esto es todo lo que hay que saber. No sé si la persona resultante será yo, o será alguien distinto pero exactamente como yo. Ahora bien, esto no es aquí una pregunta real, que tenga que tener una respuesta. No describe dos posibilidades diferentes, de las que una tenga que ser verdadera. Se trata de una cues-

[29] *The Times*, Londres, Columna de Ciencia, 22 de noviembre de 1982. Me dijeron que pronto aparecerán informes de resultados más impresionantes en las revistas científicas apropiadas.

tión vacía. No hay aquí diferencia real entre que la persona resultante sea *yo*, y que sea *alguien distinto*. Por eso lo sé todo, aunque no sepa si estoy a punto de morir.

Pienso que, para los que aceptan el criterio físico, esta es la reacción correcta a esta gama de casos. Pero la mayor parte de nosotros no aceptaría, con todo y con eso, estas afirmaciones.

Si aun así no aceptamos la concepción reduccionista, y seguimos creyendo que nuestra identidad tiene que ser determinada, ¿qué debemos decir de estos casos? Si seguimos asumiendo que mi Réplica no sería yo, nos vemos forzados a aceptar la conclusión siguiente: tiene que haber un porcentaje crítico tal que, si los cirujanos reemplazan menos que ese tanto por ciento, seré yo el que se despierte, pero si reemplazan más que este tanto por ciento, *no* seré yo, sino sólo alguien distinto que es simplemente como yo. Podríamos sugerir una variante de esta conclusión. Quizás haya una parte crucial de mi cerebro que es tal que, si los cirujanos no la sustituyen, la persona resultante será yo, pero si lo hacen, será alguien distinto. Pero esto no cambia las cosas. ¿Qué ocurre si reemplazan porcentajes diferentes de esta parte crucial de mi cerebro? De nuevo nos vemos forzados a aceptar la idea de que tiene que haber un porcentaje crítico.

Esta idea no es incoherente. Pero es difícil de creer. Y ocurre otra cosa que la hace aún más difícil de creer. No podríamos *descubrir* cual es el porcentaje crítico mediante la realización de algunos de los casos en este espectro imaginario. Yo podría decir, «Tratad de reemplazar el 50% de las células de mi cerebro y de mi cuerpo, y os diré lo que ocurre». Pero sabemos de antemano que, en cada caso, la persona resultante estaría inclinada a pensar que ella es yo. Y esto no demostraría que ella *es* yo. Realizar esos casos no aportaría la respuesta a nuestra pregunta.

Estos comentarios asumen que todos los rasgos psicológicos de una persona dependen de los estados de las células en su cerebro y en su sistema nervioso. Asumo que una Réplica orgánica mía sería exactamente como yo desde el punto de vista psicológico. Si rechazamos esta asunción, podríamos responder a esta gama de casos imaginarios de un modo diferente. Me ocupo de esta respuesta en la sección siguiente.

Si mi asunción es correcta, y cada una de estas personas resultantes fuese exactamente como yo, ¿qué deberíamos pensar de esta gama de casos? Tenemos tres alternativas:

(1) Podríamos aceptar la respuesta reduccionista que se dio arriba.
(2) Podríamos pensar que *hay* un límite divisorio claro entre dos casos. Si los cirujanos sustituyeran sólo determinadas células, la persona resultante sería yo. Si en vez de eso reemplazaran sólo unas cuantas células más, la persona resultante no sería yo, sino que simplemente sería exactamente como yo. Tiene que haber este límite divisorio claro en esta gama de casos, aunque nunca pudiéramos descubrir dónde trazar la línea.
(3) Podríamos pensar que, en todos estos casos, la persona resultante sería yo.

De estas tres conclusiones, (3) le parece a la mayoría de la gente la menos increíble. Si aceptamos (3), pensamos que la continuidad psicológica proporciona identidad personal. Pensamos que la proporciona aunque esta continuidad no tenga su causa normal: la existencia continua de un cerebro concreto.

Cuando considerábamos el Espectro Psicológico, el argumento de Williams pareció demostrar que la continuidad psicológica no es necesaria para la identidad personal. La continuidad física sería suficiente. Cuando consideramos ahora el Espectro Físico, una argumento parecido parece demostrar que la continuidad física no es necesaria para la identidad personal. La continuidad psicológica sería suficiente.

Podríamos aceptar las dos conclusiones. Podríamos sostener que cualquier tipo de continuidad asegura la identidad personal. Aunque esta concepción híbrida es coherente, está expuesta a graves objeciones. Una de ellas surge si combinamos, no nuestras dos conclusiones, sino los dos argumentos para estas conclusiones.

86. El Espectro Combinado

Consideremos otra gama de casos posibles. Implican todas las variaciones posibles en los grados de conexividad *tanto* física *como* psicológica. Se trata del *Espectro Combinado*.

En el extremo cercano de este espectro se encuentra el caso normal en que una persona futura sería del todo continua conmigo como soy ahora, tanto física como psicológicamente. Esta persona sería yo, exactamente de la misma manera que, en mi vida real, seré yo el que se despierte mañana. En el extremo lejano de este espectro, la persona resultante no tendría continuidad conmigo como soy ahora, ni física ni psicológica. En este caso, los científicos destruirían mi cerebro y mi cuerpo, y luego crearían, a partir de materia orgánica nueva, una Réplica perfecta de alguien distinto. Supongamos que esta persona no es Napoleón sino Greta Garbo. Podemos suponer que, cuando Garbo tenía 30 años, un grupo de científicos grabó los estados de todas sus células cerebrales y corporales.

En el primer caso de este espectro, en el extremo cercano, no se haría nada. En el segundo caso, unas pocas células de mi cerebro y de mi cuerpo serían reemplazadas. Las nuevas células *no* serían duplicados exactos. Como resultado, habría de alguna manera menos conexividad psicológica entre yo y la persona que se despierta. Esta persona no tendría todos mis recuerdos, y su carácter sería, de algún modo, diferente del mío. Tendría algunos recuerdos aparentes de la vida de Greta Garbo, y tendría uno de los rasgos de carácter de la Garbo. A ella le encantaría actuar, no como a mí. Su cuerpo también sería, de algún modo, menos como el mío y más como el de Garbo. Sus ojos se parecerían más a los ojos de Garbo. En una posición más distante del espectro, sería reemplazado un porcentaje mayor de mis células, también con células distintas. La persona resultante estaría conectada de menos maneras conmigo, y de más maneras con la Garbo, tal y como era a la edad de 30 años. Y habría cambios similares en el cuerpo de esa persona. Cerca del extremo más lejano, la mayoría de mis células serían reemplazadas por células distintas. La persona que se despertara tendría nada más que unas cuantas células de mi cerebro y de mi cuerpo originales, y entre ella y yo habría nada más que unas cuantas conexiones psicológicas. Tendría pocos recuerdos aparentes que encajasen con mi pasado, y pocos de mis hábitos y deseos. Pero en todos los demás aspectos la persona sería, tanto física como psicológicamente, justo como Greta Garbo.

Estos casos proporcionan, creo yo, un poderoso argumento a favor de la Concepción Reduccionista. Una vez más, el argumento asume que nuestros rasgos psicológicos dependen de los estados de nuestro cerebro. Supongamos que la causa de la continuidad psicológica no fuese la existencia continua del cerebro, sino la existencia continua de una entidad que existe separadamente, como por ejemplo un Ego Cartesiano. Podríamos afirmar entonces que, si realizáramos esas operaciones, los resultados *no* serían como los que he descrito. Encontraríamos que, si reemplazamos una gran cantidad del cerebro de alguien, incluso con células disímiles, la persona resultante sería exactamente como la persona original. Pero habría un porcentaje crítico, o una parte crítica del cerebro, cuya sustitución destruiría completamente la continuidad psicológica. En uno de los casos de esta gama, el portador de la continuidad cesaría o bien de existir, o bien de interactuar con este cerebro. La persona resultante sería, desde el punto de vista psicológico, totalmente diferente de la persona original.

Si tuviésemos razones para creer en esta concepción, ella proporcionaría una respuesta a mi argumento. *Habría*, en esta gama de casos, una línea divisoria nítida. Y *esta* línea divisoria *podría* descubrirse. Correspondería a lo que pareció ser un cambio completo de identidad personal. Y esta concepción también explicaría cómo el reemplazo de unas cuantas células podría destruir totalmente la continuidad psicológica.

Y esta concepción podría aplicarse tanto al Espectro Psicológico como al Físico. Podríamos decir que, en estos dos espectros, los resultados no serían, en efecto, los que yo supuse.

Excepto los más próximos al extremo cercano, los casos del espectro combinado son técnicamente imposibles, y es probable que seguirán siéndolo. Por eso no podemos descubrir directamente si los resultados serían como yo supuse, o serían por el contrario de la clase recién descrita. Pero cuáles serían los resultados depende de cuál sea la relación entre los estados del cerebro de alguien y la vida mental de esta persona. ¿Contamos con evidencia que nos lleve a creer que la continuidad psicológica depende principalmente, no de la continuidad del cerebro, sino de la con-

tinuidad de alguna entidad diferente, que o bien existe intacta o bien no existe en absoluto? Ciertamente no tenemos el tipo de evidencia que describí arriba. Y sí tenemos muchas razones para creer tanto que el portador de la continuidad psicológica es el cerebro, cuanto que la conexividad psicológica puede darse en cualquier grado reducido.

Como nuestros rasgos psicológicos dependen de los estados de nuestro cerebro, estos casos imaginarios son sólo técnicamente imposibles. Si pudiésemos realizar estas operaciones, los resultados serían los que he descrito. ¿Qué deberíamos pensar de los diferentes casos de este espectro combinado? ¿Cuáles son los casos en que yo seguiría existiendo?

Como antes, no podríamos dar con la respuesta llevando a cabo realmente, en mí y en otras personas, operaciones como las que hemos imaginado. Ya sabemos que, en algún lugar a lo largo del espectro, se situaría el primer caso en que la persona resultante pensaría que él o ella no era yo. Pero no tenemos razón alguna para confiar en esta creencia. En este tipo de casos, quién es alguien no puede establecerse por apelación a quién cree ser. Como los experimentos no ayudarían, tenemos que tratar de decidir ahora lo que pensamos de estos casos.

Al considerar los dos primeros espectros, nos quedamos con tres alternativas: aceptar una respuesta reduccionista, creer que tiene que haber una línea divisoria nítida, y creer que la persona resultante sería yo en cualquier caso. De las tres, la tercera pareció la conclusión menos increíble.

Al considerar el espectro combinado, no podemos aceptar esta conclusión. En el caso del extremo más lejano, los científicos destruyen mi cerebro y mi cuerpo, y después hacen una Réplica de Greta Garbo, a partir de materia nueva. No habría conexión de ninguna clase entre esta persona resultante y yo. No podría estar más claro que, en este caso, la persona resultante *no* sería yo. Nos vemos forzados a elegir entre las otras dos alternativas.

Podríamos seguir pensando que nuestra identidad tiene que ser determinada. Podríamos seguir pensando que, para la pregunta «¿La

persona resultante sería yo?», tiene que haber siempre una respuesta, que tiene que ser, simplemente, Sí o No. Entonces nos veríamos forzados a aceptar las afirmaciones siguientes:

> En algún lugar de este espectro, hay una línea divisoria nítida. Tiene que haber un conjunto crítico de células reemplazadas y un grado crítico de cambio psicológico, que supondrían toda la diferencia. Si los cirujanos reemplazan una cantidad ligeramente menor de esas células, y producen un cambio psicológico más pequeño, seré yo el que se despierte. Pero si reemplazan las pocas células de más, y producen un cambio psicológico más, yo dejaré de existir, y la persona que se despierte será alguien distinto. Tiene que haber tal par de casos en algún lugar de este espectro, *aunque no pudiera haber nunca ninguna evidencia de dónde están estos casos.*

Estas afirmaciones son difíciles de creer. Es difícil de creer (1) que la diferencia entre la vida y la muerte pueda consistir tan sólo en una de las pequeñísimas diferencias descritas arriba. Nos inclinamos a pensar que hay *siempre* una diferencia entre que una persona futura sea yo, y que sea alguien diferente. Y nos inclinamos a pensar que es una diferencia *profunda*. Pero entre los casos vecinos de este espectro las diferencias son triviales. Por eso es difícil de creer que, en uno de estos casos, la persona resultante sería yo con toda claridad, y que, en el caso siguiente, sería con toda claridad otra diferente.

Es también difícil de creer (2) que tenga que haber una tal línea divisoria nítida en algún lugar del espectro, aunque nunca fuéramos capaces de contar con ninguna evidencia de dónde estaría esa línea. Algunos dirían que si nunca va a haber tal evidencia, carece de sentido afirmar que en algún lugar tiene que encontrarse esa línea.

Aunque (2) tenga sentido, las afirmaciones (1) y (2), tomadas en conjunto, son extremadamente poco convincentes. Pienso que son incluso menos convincentes que la única posible conclusión que queda, que es la Concepción Reduccionista. Según esta concepción, en los casos centrales del Espectro Combinado, sería una

pregunta vacía la de si la persona resultante sería yo. Este espectro proporciona, como afirmé, un poderoso argumento a favor de esta concepción.

Hay quienes creen que nuestra identidad tiene que ser determinada, aunque no piensen que seamos entidades que existen separadamente, distintas de nuestros cerebros y cuerpos, y de nuestras experiencias. Pero creo que esta concepción es insostenible. ¿Qué explicaría el supuesto hecho de que la identidad personal es siempre determinada? La respuesta tiene que ser que el verdadero criterio de identidad personal cubre cualquier caso. El verdadero criterio tiene que trazar una línea divisoria nítida en algún lugar del Espectro Combinado. Pero si no somos entidades que existen separadamente, ¿cómo podría haber tal línea divisoria? ¿Qué es lo que podría hacer que en un caso la persona resultante fuese yo, y en el siguiente no lo fuese? ¿En qué podría consistir la diferencia?

Hay otras personas que creen que, aunque no seamos entidades que existen separadamente, la identidad personal es un hecho adicional. Piensan que la identidad personal no sólo consiste en las diferentes clases de continuidad física y psicológica. Esta es otra concepción que encuentro insostenible. Si no somos entidades que existen separadamente, ¿en qué podría consistir este hecho adicional? ¿Qué es lo que podría hacer, en los casos de esta gama, que este hecho se diera o no se diera?

Este espectro demuestra, pienso, que hay que mantener juntas ciertas concepciones. No podemos pensar de manera justificable que nuestra identidad implica un hecho adicional, a no ser que también pensemos que somos entidades que existen separadamente, distintas de nuestros cerebros y de nuestros cuerpos. Y no podemos pensar de manera justificable que nuestra identidad tiene que ser determinada, a no ser que creamos que la existencia de estas entidades separadas tiene que ser todo-o-nada.

Los hay que piensan que la identidad de *todo* tiene que ser siempre determinada. Estas personas aceptan una forma estricta de la doctrina de que *no hay entidad sin identidad*. Se trata de la tesis de que no podemos referirnos a un objeto en particular, ni tampoco nombrarlo, a no ser que nuestro criterio de identidad para este objeto nos brinde una respuesta definitiva en cada caso concebible. Según esta idea, con frecuencia pensamos equivocadamente que nos estamos refiriendo a un objeto, cuando, como no hay tal criterio de identidad, no hay tal objeto. Por ejemplo, se afirmaría que la mayoría de nosotros piensa equivocadamente que el nombre «Francia» refiere a una nación. Según esta concepción, no nos podemos referir a las naciones porque las naciones no existen. No hay ningún criterio de identidad para naciones que llegue al nivel requerido —que nos diga, en cada caso concebible, si una nación ha seguido o no ha seguido existiendo—. Los que mantienen esta opinión puede que piensen que no podría ser cierto, de forma parecida, que las personas no existan. Si esta opinión es verdadera, y las personas existen, el criterio de identidad personal tiene que brindar una respuesta definitiva en todos los casos.

No es necesario que esta forma de pensar conlleve la creencia de que una persona es una entidad que existe separadamente. Lo cual puede que parezca hacerla más convincente. Pero, si la hacemos nuestra, tenemos que creer otra vez que el verdadero criterio de identidad personal traza una línea divisoria nítida, del todo incognoscible, en algún lugar del Espectro Combinado. Como he dicho, si la identidad personal no conlleva un hecho adicional, es muy difícil de creer que pueda haber una línea semejante. Es aún menos plausible que la Concepción Reduccionista.

Hay otro modo en que algunos autores afirman que nuestra identidad tiene que ser determinada. Según esta concepción, tenemos creencias inconsistentes si hay casos en que no podemos responder a una pregunta sobre la identidad de un objeto. Pienso que hay casos de estos, y que en ellos la identidad de un objeto es indeterminada. Afirmo que, en un caso semejante, la afirmación «Este es el mismo objeto que teníamos antes» no sería ni verdadera ni falsa. Se ha argumentado que esta tesis es incoherente [31]. Creo que este argumento ha sido contestado [32]. Pero supongamos que es correcto.

[31] Evans (1), p. 208; véase asimismo Salmon, pp. 245-6.
[32] Véase Broome (1).

Implicaría lo siguiente: cuando damos con casos que no están cubiertos por lo que creemos que es un criterio de identidad, deberíamos revisar nuestras creencias ampliando este criterio. Si hacemos nuestra esta forma de pensar, entonces no creeremos que el verdadero criterio de identidad personal tenga que trazar una línea divisoria nítida en algún lugar del Espectro Combinado. Más bien creeremos que debemos trazar nosotros esa línea para evitar la incoherencia.

Esta concepción apenas difiere de la reduccionista. Si somos nosotros los que trazamos tal línea, no podemos creer que tenga, de una manera intrínseca, significación racional ni moral. Tenemos que seleccionar un punto de este espectro, hasta el cual llamaremos a la persona resultante yo, y más allá del cual la llamaremos alguien diferente. Nuestra elección de este punto tendrá que ser arbitraria. Tenemos que trazar esta línea entre dos casos vecinos, aunque la diferencia entre ellos sea, en sí misma, trivial. Si esto es lo que hacemos, no debería afectar a nuestra actitud hacia estos dos casos. Sería para mí claramente irracional considerar el primer caso tan bueno como la supervivencia corriente, y al segundo tan malo como la muerte corriente. Cuando considero esta gama de casos, pregunto naturalmente, «¿La persona resultante será yo?». Al trazar nuestra línea, hemos elegido *dar* una respuesta a esta pregunta. Pero, como nuestra elección fue arbitraria, no puede justificar cualquier tesis sobre lo que importa. Si respondemos así a la pregunta sobre mi identidad, hemos hecho que, en esta gama de casos, la identidad personal *no* sea lo que importa. Y esta es la tesis más importante de la Concepción Reduccionista. Nuestra concepción difiere sólo trivialmente de esta concepción. Los reduccionistas afirman que, en algunos casos, las preguntas sobre la identidad personal son indeterminadas. Nosotros añadimos la tesis de que, en tales casos, debemos dar respuesta a estas preguntas, aunque tengamos que hacerlo de forma arbitraria, una forma que despoja a nuestras respuestas de toda significación. Considero a esta concepción una versión del reduccionismo, la versión propia de una mente ordenada que elimina la indeterminación con definiciones convencionales faltas de interés. Como la diferencia es tan leve, la ignoraré.

Según la versión más simple del Fisicalismo, todo suceso mental es un suceso que tiene lugar en un cerebro. Subrayé arriba que podemos ser fisicalistas y al mismo tiempo aceptar el Criterio *Psicológico* de identidad personal. Debo añadir que no hace falta que los reduccionistas sean fisicalistas. Y si no somos fisicalistas, podemos ser o dualistas, que creen que los sucesos mentales son diferentes de los sucesos físicos, o idealistas, que creen que todos los sucesos son puramente mentales. Si pensamos que somos Egos Cartesianos, creemos en una forma de dualismo. Pero los dualistas pueden ser reduccionistas en el asunto de la identidad personal. Podemos creer que los sucesos mentales son distintos de los sucesos físicos, y creer además que la unidad de la vida de una persona no consiste más que en las diversas clases de conexión que se dan entre todos los sucesos mentales y físicos que, juntos, constituyen su vida. Esta sería la versión dualista de la concepción reduccionista.

Defenderé que, si somos reduccionistas, no deberíamos tratar de decidirnos entre los diferentes criterios de identidad personal. Una razón es que la identidad personal no es lo que importa. Antes de que defienda esta conclusión, explicaré con más detalle lo que afirma un reduccionista. Y como la mayoría de nosotros se halla fuertemente inclinada a rechazar estas afirmaciones, considerar el Espectro Combinado puede que no sea suficiente para cambiar de forma de pensar. Por tanto plantearé, en el próximo capítulo, otros argumentos a favor de la Concepción Reduccionista.

Los reduccionistas admiten que hay una diferencia entre la identidad numérica y la similitud exacta. En algunos casos, habría una diferencia real entre que una persona sea yo y que sea alguien distinto que, simplemente, es exactamente como yo. Muchos dan por sentado que *siempre* tiene que haber tal diferencia.

En el caso de las naciones, o en el de los clubes, semejante asunción es falsa. Dos clubes podrían existir al mismo tiempo, y ser, aparte de sus miembros, exactamente iguales. Si soy miembro de uno de esos clubes, y tú dices que también eres miembro, yo podría preguntar «¿Eres miembro de exactamente el mismo club

del que yo soy miembro? ¿O eres simplemente miembro del otro club, que es exactamente igual?». No es esta una pregunta vacía, pues describe dos posibilidades diferentes. Pero aunque haya dos posibilidades en el caso en que los dos clubes coexisten, puede que no haya estas dos posibilidades cuando discutimos la relación entre un club que existe en el presente y un club pasado. No había dos posibilidades en el caso que describí en la sección 79. En este caso no había nada que justificase la afirmación de que tenemos exactamente el mismo club ni la de que tenemos un club nuevo que es, simplemente, exactamente igual. En ese caso *no* se trataría de dos posibilidades diferentes.

Del mismo modo, hay algunos casos en que hay una diferencia real entre que alguien sea yo y que sea alguien distinto que es exactamente como yo. Puede ocurrir así en el caso de la línea secundaria, la versión del teletransporte en que el escáner no destruye mi cerebro ni mi cuerpo. En el caso de la línea secundaria, mi vida se superpone parcialmente con la vida de mi Réplica en Marte. Dada esta superposición, podemos concluir que somos dos personas diferentes —que somos cualitativa, pero no numéricamente idénticas—. Si soy la persona que queda en la Tierra, y ahora existe mi Réplica en Marte, supone una diferencia el que un dolor sea sentido por mí, o sea sentido en cambio por mi Réplica. Esta es una diferencia real en lo que ocurre.

Si volvemos al teletransporte simple, donde no hay superposición parcial entre mi vida y la de mi Réplica, las cosas son diferentes. Aquí podríamos decir que mi Réplica será yo, o podríamos decir en vez de eso que será simplemente alguien distinto que es exactamente como yo. Pero no deberíamos considerar que se trata de hipótesis rivales acerca de lo que va a suceder. Para que se trate de hipótesis rivales, mi existencia continua tiene que implicar un *hecho adicional*. Si mi existencia continua implica simplemente continuidad física y psicológica, sabemos exactamente lo que ocurre en este caso. Habrá una persona futura que será exactamente como yo desde el punto de vista físico, y que será totalmente continua conmigo desde el punto de vista psicológico. Esta continuidad psicológica tendrá una causa fiable, la transmisión de mi cianotipo. Pero

esta continuidad no tendrá su causa normal, puesto que esta persona futura no será físicamente continua conmigo. Esta es una descripción completa de los hechos. No hay ningún hecho adicional del que seamos ignorantes. Si la identidad personal no implica un hecho adicional, no deberíamos creer que aquí hay dos posibilidades diferentes: que mi Réplica será yo o que será alguien distinto que es simplemente como yo. ¿Qué podría hacer a éstas posibilidades diferentes? ¿En qué podría consistir la diferencia?

Algunos no-reduccionistas estarían de acuerdo en que, en este caso, no hay dos posibilidades. Piensan que, en el caso del teletransporte, mi Réplica no sería yo. Después discutiré un argumento plausible a favor de esta conclusión. Si estuviéramos equivocados al decir que mi Réplica es yo, los comentarios que he hecho ahora mismo se aplicarían en cambio a los casos centrales del Espectro Físico. Mi Réplica podría tener un cuarto de las células existentes en mi cerebro y mi cuerpo, o la mitad, o tres cuartos. En estos casos no hay dos posibilidades diferentes: que mi Réplica sea yo, o que sea alguien distinto que es simplemente como yo. Se trata de descripciones simplemente diferentes del mismo resultado.

Si pensamos que hay siempre una diferencia real entre que una persona sea yo y que sea alguien distinto, tenemos que creer que esta diferencia llega en algún lugar de esta gama de casos. Tiene que haber una línea divisoria nítida, aunque nunca pudiéramos saber dónde está. Como he dicho, esta creencia aún es menos convincente que la Concepción Reduccionista.

En el caso de los clubes, aunque a veces haya una diferencia entre la identidad numérica y la similitud exacta, a veces no hay tal diferencia. La pregunta, «¿Es el mismo, o, simplemente, exactamente igual?» es a veces vacía. Esto también podría ser verdadero de las personas. Sería verdadero o al final o en el medio del Espectro Físico.

Es difícil de creer que esto pudiera ser verdadero. Cuando me imagino a mí mismo a punto de apretar el botón verde, es difícil de creer que no haya una pregunta real relativa a si estoy a punto de

morir, o en cambio despertaré de nuevo en Marte. Pero, como he defendido, esta creencia no puede justificarse a no ser que la identidad personal implique un hecho adicional. Y no podría darse tal hecho a no ser que yo sea una entidad que existe separadamente, aparte de mi cerebro y de mi cuerpo. Una entidad tal es un Ego Cartesiano. Como he afirmado, no hay ninguna evidencia a favor de esta tesis, y sí mucha evidencia en contra de ella.

12

POR QUÉ NUESTRA IDENTIDAD NO ES LO QUE IMPORTA

87. MENTES DIVIDIDAS

Algunos casos médicos recientes proporcionan una evidencia impresionante a favor de la Concepción Reduccionista. Los seres humanos tenemos un cerebro inferior y dos hemisferios superiores que están conectados por un haz de fibras. Para tratar a unos cuantos pacientes con epilepsia severa, los cirujanos han cortado esas fibras. El propósito no era otro que reducir la gravedad de los ataques epilépticos confinando sus causas a un solo hemisferio. Se logró lo que se pretendía, pero las operaciones tuvieron otra consecuencia no deseada. El efecto, en palabras de un cirujano, era la creación de «dos esferas separadas de conciencia» [33]:

Diversas pruebas psicológicas pusieron de manifiesto este efecto. Las pruebas hacían uso de dos hechos. Controlamos el brazo derecho con el hemisferio izquierdo, y viceversa. Y vemos lo que se encuentra en la mitad derecha del campo visual con nuestro hemisferio izquierdo, y viceversa. Cuando los hemisferios de un individuo han sido desconectados, los psicólogos pueden, por ejemplo, pre-

[33] Sperry, p. 299.

sentarle dos preguntas diferentes, escritas en las dos mitades de su campo visual, y pueden recibir dos respuestas diferentes, escritas por cada una de las manos de esta persona.

Aquí tenemos una versión simplificada de la clase de evidencia que proporcionan estas pruebas: a una de estas personas se le muestra una pantalla ancha, su mitad izquierda es roja, y su mitad derecha es azul. En cada mitad, con un tono más oscuro, están escritas las palabras, «¿Cuántos colores ves?». Con cada mano la persona escribe, «Solo uno». Entonces se cambian las palabras, de manera que se lee, «¿Cuál es el color que ves?». Con una de las manos la persona escribe «Rojo», y con la otra escribe «Azul».

Si la persona responde así, no parece que haya razón para dudar de que está teniendo sensaciones visuales —que ve, tal y como afirma, rojo y azul—. Pero al ver rojo no es consciente de ver azul, y viceversa. Por eso el cirujano escribe acerca de «dos esferas de conciencia separadas». En cada uno de sus centros de conciencia la persona sólo puede ver un color. En un centro, ve rojo, en el otro, azul.

Las muchas pruebas reales que se han hecho muestran estos dos mismos rasgos esenciales, aunque difieren, en los detalles, de la prueba imaginaria que acabo de describir. Al ver lo que se encuentra en la mitad izquierda de su campo visual, la persona no es en absoluto consciente de lo que está viendo en ese mismo momento en la mitad derecha de su campo visual, y viceversa. Y en el centro de conciencia en que ve la mitad izquierda de su campo visual, y en que es consciente de lo que está haciendo con la mano izquierda, la persona es completamente inconsciente de lo que está haciendo con la mano derecha, y viceversa.

Una de las complicaciones de los casos reales es que en la mayoría de las personas, por lo menos en las primeras semanas después de la operación, el lenguaje está enteramente controlado por el hemisferio que lleva la mano derecha. Como resultado, «si la palabra "sombrero" se ilumina en el lado izquierdo, la mano izquierda elegirá un sombrero de un grupo de objetos ocultos si a la persona se le dijo que cogiera lo que vio. Al mismo tiempo, insistirá verbalmente en que no vio nada» [34]. Otra complicación es que, tras

[34] Nagel (5), reeditado en Nagel (4), p. 152.

cierto tiempo, cada hemisferio puede a veces controlar las dos manos. Nagel cita un ejemplo de la clase de conflicto que puede darse entonces:

> «Se coloca una pipa [pipe] fuera de la vista en la mano izquierda del paciente, después se le pide que escriba con la mano izquierda el nombre del objeto que estuvo sosteniendo. Con gran esfuerzo y con trazo muy marcado, la mano izquierda escribe las letras P e I. Luego de repente la escritura se hace más rápida y ligera, la I se convierte en E, y la palabra se completa como LÁPIZ [pencil]. Evidentemente, el hemisferio izquierdo ha hecho una suposición basada en la apariencia de las dos primeras letras, y ha interferido... Pero entonces el hemisferio derecho recobra el control de la mano, tacha con trazo muy marcado las letras ENCIL, y dibuja una tosca imagen de una pipa» [35].

Un conflicto semejante puede adoptar formas más siniestras. Uno de los pacientes se quejaba de que a veces, cuando abrazaba a su mujer, la mano izquierda la apartaba de un empujón.

Mucho se ha dicho de otra complicación que se presenta en los casos reales, complicación a la que el ejemplo de Nagel hace alusión. El hemisferio izquierdo generalmente soporta o «tiene» las capacidades lingüísticas y matemáticas de un adulto, mientras que el hemisferio derecho «tiene» esas capacidades al nivel de un niño pequeño. Pero el hemisferio derecho, aunque menos adelantado en estos aspectos, tiene más destrezas de otro tipo, como las que están implicadas en el reconocimiento de patrones, o en el sentido musical. Se asume que, después de cumplir los tres o cuatro años, los dos hemisferios siguen una «división del trabajo», desarrollando cada uno determinadas capacidades. Que el hemisferio derecho tenga una menor capacidad lingüística no es algo intrínseco ni permanente. Las personas que han padecido derrames en el hemisferio izquierdo con frecuencia retornan a la capacidad lingüística de un niño pequeño, pero con el hemisferio derecho que les queda sano muchos pueden reaprender el habla adulta. Se cree también que en una minoría

[35] Nagel, *op. cit.*, p. 153.

de personas puede que no haya diferencia entre las destrezas de los dos hemisferios.

Supongamos que pertenezco a esa minoría, con dos hemisferios exactamente iguales. Y supongamos que he sido equipado con algún dispositivo que puede bloquear la comunicación entre ambos. Como el dispositivo se halla conectado a mis cejas, está bajo mi control. Levantando una ceja puedo dividir mi mente. En cada mitad de mi mente dividida puedo después, bajando una ceja, reunificar mi mente.

Esta destreza podría tener mucha utilidad. Consideremos

> *Mi Examen de Física.* Estoy en un examen, y me quedan sólo quince minutos para responder a la última pregunta. Se me ocurre que hay dos maneras de abordarla, pero no estoy seguro de cuál es más probable que tenga éxito. Por eso tomo la decisión de dividir mi mente durante diez minutos, para que cada mitad de ella trabaje en uno de los dos cálculos, y después reunificarla para redactar una copia en limpio del mejor resultado. ¿Cómo será la experiencia?

Cuando desconecto mis hemisferios, mi corriente de conciencia se divide. Pero esta división no es algo que yo experimente. Cada una de mis dos corrientes de conciencia parece haber sido directamente continua con mi única corriente de conciencia hasta el momento de la división. Los únicos cambios en cada corriente son la desaparición de la mitad del campo visual y la pérdida de la sensación y del control en uno de los brazos.

Consideremos mis experiencias en la corriente «que controla la mano derecha». Recuerdo que decidí usar la mano derecha para hacer el cálculo más largo. Este es el cálculo que ahora comienzo. Al trabajar en él puedo ver, por el movimiento de la mano izquierda, que también estoy trabajando en el otro. Pero no soy consciente de trabajar en el otro. En la corriente que lleva la mano derecha podría preguntarme cómo me va en la corriente que controla la mano izquierda. Podría mirar y ver. Lo cual sería justo como mirar a ver cómo lo está haciendo mi vecino, en el pupitre de al lado. En la corriente que controla la mano derecha yo sería tan inconsciente de

lo que está pensando ahora mi vecino como de lo que yo estoy pensando ahora en la corriente que controla la mano izquierda. Las mismas observaciones se aplican a mis experiencias en la corriente que controla la mano izquierda.

Acabo de terminar mi trabajo. Estoy a punto de reunificar mi mente. ¿Qué debo esperar que ocurra en cada corriente? Sólo que de repente me dará la impresión de recordar haber trabajado en dos cálculos, y que trabajando en cada uno de los dos no fui consciente de trabajar en el otro. Esto es lo que sugiero que podemos imaginar. Y, si mi mente hubiera sido dividida, mis recuerdos aparentes serían correctos.

Al describir este caso, di por sentado que había dos series separadas de pensamientos y sensaciones. Si se pudo ver que mis manos escribieron dos cálculos, y también dije recordar después dos series correspondientes de pensamientos, esto es lo que debemos dar por sentado. Sería completamente increíble asumir que uno o los dos cálculos habían sido hechos de manera inconsciente.

Podría objetarse que mi descripción ignora «la necesaria unidad de la conciencia». Pero no he ignorado esta supuesta necesidad. Lo que he hecho es negarla. Lo que es un hecho tiene que ser posible. Y es un hecho que las personas con los hemisferios desconectados tienen dos corrientes de conciencia separadas —dos series de pensamientos y de experiencias, de forma que al tener una son inconscientes de tener otra—. Cada una de estas dos corrientes exhibe por separado unidad de conciencia. Lo cual puede ser un hecho sorprendente. Pero podemos entenderlo. Podemos llegar a pensar que la historia mental de una persona no tiene por qué ser como un canal con un solo cauce, sino que podría ser como un río que de vez en cuando tiene corrientes separadas. Sugiero que también podemos imaginar qué se sentiría al dividir y reunificar nuestras mentes. Mi descripción de mis experiencias en mi examen de Física parece ser coherente, y además describir algo que nos podemos imaginar.

Podría decirse además que, en mi caso imaginario, no tengo una mente dividida. Más bien tengo dos mentes. Esta objeción no plan-

tea una cuestión real, porque se trata de dos modos de describir el mismo resultado.

Una objeción parecida afirma que, en estos casos reales e imaginarios, el resultado no es una persona con una mente dividida o con dos mentes. El resultado son dos personas diferentes, compartiendo el control de la mayor parte de un mismo cuerpo, pero cada una de ellas controlando en exclusiva un brazo. Tampoco aquí pienso que esto plantee una cuestión real. Se trata una vez más de dos modos de describir el mismo resultado. Esto es lo que pensaremos si somos reduccionistas.

Si todavía no somos reduccionistas, que es lo que doy por supuesto, pensaremos que es una cuestión real si estos casos implican a más de una persona. Quizás podamos creerlo así en los casos reales, en los que la división es permanente. Pero esta creencia es difícil de aceptar cuando consideramos mi imaginario Examen de Física, porque en él hay dos corrientes de conciencia durante sólo diez minutos. Y a mí después me da la impresión de recordar hacer los dos cálculos que, durante esos diez minutos, se pudo ver a mis manos escribiendo. Dada la naturaleza modesta y breve de esta falta de unidad, no resulta convincente afirmar que este caso implica a más de una persona. ¿Vamos a suponer que, durante esos diez minutos, yo dejo de existir y dos personas nuevas vienen a la existencia, y que entonces cada una de ellas se pone a trabajar en los cálculos? Según esta interpretación, todo el episodio implica a tres personas, dos de las cuales tienen vidas que duran sólo diez minutos. Además, cada una de estas dos personas piensa erróneamente que ella es yo, y tiene recuerdos aparentes que se corresponden fielmente con mi pasado. Y después de estos diez minutos yo tengo recuerdos aparentes fieles de las breves vidas de cada una de estas dos personas, salvo que erróneamente me figuro que yo mismo tuve todos los pensamientos y las sensaciones que tuvieron estas personas. Es difícil de creer que me equivoco aquí, y que el episodio implica a tres personas totalmente diferentes.

Es igual de difícil pensar que implica a dos personas diferentes, a mí haciendo uno de los cálculos y a otra persona haciendo el otro. Admito que, cuando divido mi mente por primera vez, yo podría

pensar al hacer uno de los cálculos que el otro cálculo tenía que estar siendo hecho por otro. Pero al hacer el otro cálculo podría tener la misma creencia. Y cuando mi mente hubiera sido reunificada me daría la impresión de recordar haber pensado, mientras hacía cada uno de los cálculos, que el otro cálculo tenía que estar siendo hecho por alguien diferente. Cuando me diese la impresión de recordar ambas creencias no tendría razones para pensar que una era verdadera y la otra falsa. Y, tras varias divisiones y reunificaciones, yo dejaría de tener tales creencias. En cada una de mis dos corrientes de conciencia pensaría que, en mi otra corriente, estaba en ese momento teniendo pensamientos y sensaciones de las que, en esta corriente, era en ese momento inconsciente.

88. ¿Qué es lo que explica la unidad de la conciencia?

Supongamos que, como aún no somos reduccionistas, estamos convencidos de que tiene que haber una genuina respuesta a la pregunta, «¿Quién tiene cada corriente de conciencia?». Y supongamos que, por las razones que acabo de dar, creemos que este caso implica sólo a una persona: yo. Lo que creemos es que durante diez minutos tengo una mente dividida.

Recordemos a renglón seguido la tesis de que la unidad psicológica se explica por la propiedad. Según esta tesis, debemos explicar la unidad de conciencia de una persona, en un momento dado, adscribiéndole diferentes experiencias a esa persona, en calidad de «sujeto de experiencias». Lo que da unidad a estas diferentes experiencias es que están siendo tenidas por la misma persona. Esta tesis la mantienen tanto los que creen que una persona es una entidad que existe separadamente cuanto algunos de los que rechazan esta creencia. Y esta tesis también se aplica a la unidad de cada vida.

Cuando consideramos mi imaginario Examen de Física, ¿podemos seguir aceptándola? Creemos que, mientras mi mente está dividida, tengo dos series separadas de experiencias, de forma que al tener cada una de ellas soy inconsciente de tener la otra. En un momento dado, en una de mis corrientes de conciencia, estoy

teniendo varios pensamientos y sensaciones diferentes. Podría ser consciente de meditar a fondo alguna parte del cálculo, de sentir el calambre del escritor en una mano, y de oír el ruido que hace al escribir el anticuado bolígrafo de mi vecino. ¿Qué es lo que unifica estas diferentes experiencias?

Según la tesis descrita arriba, la respuesta es que estas son las experiencias que son tenidas por mí en este momento. Esta respuesta es incorrecta. No estoy teniendo sólo estas experiencias en este momento. También estoy teniendo, en mi otra corriente de conciencia, otras varias. Necesitamos explicar la unidad de la conciencia en el interior de cada una de mis dos corrientes de conciencia, o en cada mitad de mi mente dividida. No podemos explicar estas dos unidades afirmando que todas esas experiencias están siendo tenidas por mí en este momento. Esto hace de las dos unidades una, pasando por alto el hecho de que, al tener cada uno de estos dos conjuntos de experiencias, soy inconsciente de tener el otro.

Supongamos que seguimos creyendo que la unidad debe explicarse adscribiendo diferentes experiencias a un único sujeto. Entonces tenemos que pensar que este caso implica como mínimo a dos sujetos de experiencias. Lo que unifica las experiencias en la corriente que controla mi mano izquierda es que todas ellas están siendo tenidas por un sujeto de experiencias. Lo que unifica las experiencias en la corriente que controla mi mano derecha es que todas ellas están siendo tenidas por otro sujeto de experiencias. Ahora tenemos que abandonar la afirmación de que el «sujeto de experiencias» es la persona. Según nuestra concepción, yo soy un sujeto de experiencias. Mientras mi mente está dividida hay dos sujetos de experiencias diferentes. No son el mismo sujeto de experiencias, de modo que los dos no pueden ser yo. Como es inverosímil que yo sea uno de los dos, dada la similitud de mis dos corrientes de conciencia, deberíamos concluir probablemente que yo no soy ninguno de estos dos sujetos de experiencias. Por consiguiente el episodio completo implica a tres de tales entidades. Y dos de ellas no puede decirse que sean de la clase de entidad con la que todos estamos familiarizados, una persona. Yo soy la única persona impli-

cada, y dos de estos sujetos de experiencia *no* son yo. Aunque asumamos que *soy* uno de esos sujetos de experiencia, *el otro* no puede ser yo, y por eso no es una persona.

Ahora podemos ser escépticos. Mientras el «sujeto de experiencias» era la persona, parecía verosímil afirmar que lo que unifica un conjunto de experiencias es que son todas tenidas por un único sujeto. Si tenemos que creer en sujetos de experiencias que no son personas, podemos poner en duda que haya realmente tales cosas. En el mundo animal por supuesto que hay muchos sujetos de experiencias que no son personas. Mi gato es un ejemplo. Pero otros animales son irrelevantes para este caso imaginario. Según la tesis descrita arriba, tenemos que creer que la vida de una *persona* podría implicar a sujetos de experiencias que no sean personas.

Reconsideremos mis experiencias en la corriente de conciencia que controla mi mano derecha. En esta corriente, en un momento dado, soy consciente de meditar sobre una parte de un cálculo, de sentir el calambre del escritor y de oír los ruidos que hace el bolígrafo de mi vecino. ¿Acaso explicamos la unidad de estas experiencias afirmando que todas ellas están siendo tenidas por el mismo sujeto de experiencias, y que este es una entidad que *no* soy yo? Esta explicación no parece verosímil. Si este sujeto de experiencias *no* es una persona, ¿qué clase de cosa es? No podemos decir que un Ego Cartesiano, si se dice que yo soy un Ego semejante. Este sujeto de experiencias no puede decirse que sea un Ego semejante porque no es yo, y este caso implica sólo a una persona. ¿Puede ser este sujeto de experiencias un sub-ego cartesiano, una entidad puramente mental que es simplemente parte de una persona? Podemos decidir que no tenemos razones suficientes para creer que tales cosas existan.

Y ahora me vuelvo a la otra concepción mencionada antes. Hay quienes creen que la unidad se explica por la propiedad, aunque niegan que seamos entidades que existen separadamente. Opinan que lo que unifica las experiencias de una persona en un momento dado es el hecho de que estas experiencias están siendo tenidas por esa persona. Como hemos visto, en este caso imaginario esta creencia es falsa. Mientras estoy teniendo un conjunto de experiencias en la corriente que controla mi mano derecha, también estoy teniendo

otro conjunto en la corriente que controla mi mano izquierda. No podemos explicar la unidad de ningún conjunto de experiencias afirmando que son las experiencias que estoy teniendo yo en este momento, puesto que esto combinaría los dos conjuntos.

Ahora puede intervenir el reduccionista. Según su concepción, lo que unifica mis experiencias en la corriente que controla mi mano derecha es que hay, en un momento dado, un único estado de conciencia de estas diversas experiencias. Hay un estado de conciencia de tener determinados pensamientos, de sentir el calambre del escritor y de oír el ruido de un bolígrafo escribiendo. Al mismo tiempo, hay otro estado de conciencia de las diversas experiencias en la corriente que controla mi mano izquierda. Mi mente está dividida porque no hay un único estado de conciencia de estos dos conjuntos de experiencias.

Puede objetarse que estas afirmaciones no explican la unidad de conciencia en cada corriente, sino que sólo la redescriben. Hay un sentido en que esto es cierto. Esta unidad no requiere una explicación profunda. Es simplemente un hecho el que varias experiencias puedan ser *co-conscientes*, o sea, el que sean los objetos de un único estado de conciencia. Puede ayudar la comparación de este hecho con el de que hay memoria a corto plazo de las experiencias que han tenido lugar en los últimos momentos: memoria a corto plazo de lo que se llama «el presente especioso». Igual que puede haber un recuerdo de acabar de tener varias experiencias, como oír tres veces el tañido de una campaña, puede haber un único estado de conciencia tanto de oír el cuarto tañido de la campana como de ver a los cuervos pasar volando junto a la torre de la campana. Los reduccionistas mantienen que no hay nada más que esté implicado en la unidad de conciencia en un momento determinado. Como puede haber un estado de conciencia de varias experiencias, no necesitamos explicar esta unidad adscribiendo estas experiencias a la misma persona, o sujeto de experiencias.

Vale la pena reformular otras partes de la Concepción Reduccionista. Afirmo que:

> Como adscribimos pensamientos a pensadores, es cierto que existen los pensadores. Pero los pensadores no son entidades que existan separadamente. La existencia de un pensador nada más que conlleva la existencia de su cerebro y de su cuerpo, la realización de sus actos, el pensar de sus pensamientos, y la ocurrencia de ciertos otros sucesos físicos y mentales. Podríamos, por consiguiente, redescribir la vida de una persona en términos impersonales. Al explicar la unidad de esta vida, no tendríamos necesidad de afirmar que es la vida de una persona concreta. Podríamos describir lo que, en momentos diferentes, se pensó, se sintió, se observó y se hizo, y la manera en que estos diversos sucesos estuvieron interrelacionados. Las personas sólo se mencionarían aquí en las descripciones del contenido de muchos pensamientos, deseos, recuerdos, etc. No tenemos necesidad de afirmar que las personas son los sujetos pensantes de cualquiera de estos pensamientos.

El caso en que divido mi mente apoya estas afirmaciones. Aquí no es sólo cierto que la unidad de diferentes experiencias no *necesite* explicarse adscribiéndome a mí la totalidad de esas experiencias. La unidad de mis experiencias en cada una de las corrientes *no puede* explicarse así. Hay sólo dos alternativas. Podríamos adscribir las experiencias en cada corriente a un sujeto de experiencias que *no* soy yo, y que, por tanto, no es una persona. O bien, si dudamos de la existencia de tales entidades, podemos aceptar la explicación reduccionista. Al menos en este caso, esta puede parecer ahora la mejor explicación.

Este es uno de los puntos en que importa si mi caso imaginario es posible. Si pudiésemos dividir nuestra mente por un breve espacio de tiempo, se pondría en tela de juicio la idea de que la unidad psicológica se explica por la propiedad. Como defendí, si no somos reduccionistas debemos considerar que mi caso imaginario implica nada más que a una persona. Entonces se hace posible afirmar que la unidad de conciencia debe explicarse adscribiendo diferentes experiencias a un único sujeto, la persona. Podríamos mantener esta tesis sólo si creemos en sujetos de experiencias que no son personas. Aquí son irrelevantes los otros animales. Nuestra creencia es sobre lo que la vida de las personas conlleva. Si tenemos que admitir que en estas vidas podría haber dos clases de sujetos de experiencias, los que son y los que no son personas, nuestra concepción habrá perdido gran parte de su verosimilitud. Ayudaría a

nuestra concepción el poder afirmar que, como las personas son indivisibles, mi caso imaginario nunca podría ocurrir.

Mi caso es imaginario. Pero su rasgo esencial, la división de la conciencia en corrientes separadas, *ha ocurrido* varias veces. Esto arruina la respuesta que se acaba de dar. Mi caso imaginario puede muy bien hacerse posible, y podría como mucho ser nada más que técnicamente imposible. Y en este caso la unidad de la conciencia en cada corriente no puede explicarse adscribiéndome a mí mis experiencias. Como mi explicación falla, este caso refuta la idea de que la unidad psicológica puede explicarse adscribiendo diferentes experiencias a una única persona.

Según la versión mejor conocida de esta concepción, nosotros somos Egos Cartesianos. Defendí la objeción de Lichtenberg a esta concepción cartesiana. Pero esa defensa sólo demostró que no podríamos deducir de la naturaleza de nuestras experiencias que seamos semejantes entidades. Después afirmé que no hay evidencia a favor de esta concepción, y sí mucha evidencia contra ella. Como apoyan el argumento que acabo de dar, los casos reales de mentes divididas constituyen una evidencia adicional contra esta concepción.

La idea de Descartes la podemos comparar con la creencia de Newton en el espacio y el tiempo absolutos. Newton pensaba que cualquier suceso físico tenía su posición particular sólo en virtud de su relación con estas dos realidades independientes, espacio y tiempo. Ahora pensamos que un suceso físico tiene la posición espaciotemporal que tiene en virtud de sus diversas relaciones con los otros sucesos físicos que tienen lugar. Según la Concepción Cartesiana, un suceso mental concreto ocurre dentro de una vida particular exclusivamente en virtud de su adscripción a un Ego particular. Podemos negar por nuestra parte que la topografía del «espacio mental» venga dada por la existencia de tales Egos persistentes. Podemos afirmar que un suceso mental concreto ocurre dentro de una vida en virtud de sus relaciones con muchos otros sucesos mentales y físicos, los cuales, por estar interrelacionados, constituyen esta vida [36].

[36] Madell, p. 137, sugiere que lo que hace mías a mis experiencias no es que sean tenidas por un sujeto de experiencias particular, yo, sino que tienen la *propie-*

Algunas veces se da otra razón para creer en estos Egos. Puede sostenerse, contra toda descripción completamente *objetiva* de la realidad —toda descripción que no esté hecha desde un «punto de vista»— que hay ciertas verdades que omite. Un ejemplo de tales verdades es que yo soy yo, o que yo soy Derek Parfit. Yo soy *esta* persona particular. Estas verdades *subjetivas* puede parecer que implican el que seamos sujetos de experiencias que existen separadamente.

Pero esas verdades las puede formular el reduccionista. La palabra «subjetivo» es engañosa. Las que se llaman verdades subjetivas no hace falta que impliquen a ningún sujeto de experiencias. Un pensamiento concreto puede *referirse a sí mismo*. Puede ser el pensamiento de que este pensamiento concreto, aunque exactamente igual que otros pensamientos que son pensados, es todavía *este* pensamiento concreto —o este pensar concreto de este pensamiento—. Este pensamiento es una verdad impersonal, pero subjetiva.

Algunos objetarían que todos los otros conceptos *indéxicos* —tales como «aquí», «ahora», y «este»— tienen que explicarse de un modo que utilice el concepto «yo». Pero no es así. Todos los

dad de ser mías. Según esta concepción, la topografía del espacio mental está dada por la existencia de un número muy elevado de propiedades diferentes, una para cada persona que vive alguna vez. Coincido con Madell en que yo y él podríamos tener dos experiencias simultáneas que fueran cualitativamente idénticas pero sencillamente distintas. Pero esto no tiene que ser porque una de las experiencias tenga la propiedad única de ser mía, y la otra la propiedad única de ser de Madell. Podría ser simplemente porque una de estas experiencias es *esta* experiencia, ocurriendo en *esta* vida mental particular, y la otra es *esa* experiencia, ocurriendo en *esa* otra vida mental particular. Se podría haber hecho referencia públicamente a estas dos vidas mentales a través de sus conexiones con un par de cuerpos humanos diferentes. (Los casos reales de mentes divididas proporcionan una objeción a la Concepción de Madell. Y no parece plausible tratar *ser mía* como una *propiedad*. Lo que distingue a diferentes cosas o eventos particulares no es que cada uno tenga una propiedad única. Las cosas o eventos físicos pueden ser diferentes al estar en lugares diferentes. Cuando pienso en mi pensamiento presente como siendo mío, no lo identifico por referencia a su localización espacial. Mi referencia identificadora conlleva esencialmente una palabra *indéxica*, o un *demostrativo*, antes que la adscripción de una propiedad única. Puedo usar el indéxico «este» o el indéxico «mío». Mi afirmación es que, como puedo echar mano del uso auto-referencial de «este», no necesito usar «mío».)

demás, incluyendo a «yo», pueden explicarse de un modo que emplee el uso auto-referencial de «este». Y este uso *auto*-referencial no implica la noción de un yo, o sujeto de experiencias. Con este uso de «este» podemos expresar verdades «subjetivas» sin tener que creer en la existencia separada de sujetos de experiencias [37].

89. ¿QUÉ ES LO QUE OCURRE CUANDO ME DIVIDO?

Ahora pasaré a describir otra prolongación normal de los casos reales de mentes divididas. Supongamos en primer lugar que soy uno de dos gemelos idénticos, y que tanto mi cuerpo como el cerebro de mi gemelo han resultado fatalmente lesionados. Gracias a los avances de la neurocirugía, no es inevitable que estas lesiones nos causen a los dos la muerte. Entre los dos tenemos un cerebro sano y un cuerpo sano. Los cirujanos los pueden poner juntos.

Esto podría hacerse incluso con las técnicas existentes. De la misma manera que podrían extraerme el cerebro y mantenerlo vivo conectándolo a una máquina que haga las veces de corazón y de pulmones artificiales, podría ser mantenido vivo por una conexión con el corazón y los pulmones del cuerpo de mi gemelo. Lo que nos echa atrás hoy es que los nervios de mi cerebro no podrían conectarse a los nervios del cuerpo de mi gemelo. De manera que mi cerebro podría sobrevivir si lo trasplantáramos a su cuerpo, pero la persona resultante sería paralítica.

Aunque fuera paralítica, la persona resultante podría ser habilitada para comunicarse con los demás. Un método tosco sería un dispositivo conectado con el nervio que tuviera el control del pulgar derecho de la persona, y que le permitiera enviar mensajes en código Morse. Otro dispositivo, conectado a un nervio sensorial, le podría poner en disposición de recibir mensajes. A muchas personas les gustaría sobrevivir, incluso totalmente paralizadas, con tal de

[37] Hay quienes niegan que «yo» pueda ser explicado recurriendo al uso auto-referencial de «éste». Para un argumento que apoya el punto de vista que yo acepto, véase Russell.

poder comunicarse con los demás. El ejemplo típico es el de un gran científico cuyo fin principal en la vida es seguir pensando sobre ciertos problemas abstractos.

Supongamos que, sin embargo, los cirujanos pudieran conectar mi cerebro a los nervios del cuerpo de mi gemelo. La persona resultante no tendría parálisis alguna y sería completamente sana. ¿Quién sería esa persona?

No es una pregunta difícil. Puede parecer que hay aquí un desacuerdo entre los criterios físico y psicológico. Aunque la persona resultante sería psicológicamente continua conmigo, no tendría todo mi cuerpo. Pero, como ya he dicho, el criterio físico no debe requerir la existencia continua de la totalidad de mi cuerpo.

Si todo mi cerebro sigue existiendo y sigue siendo el cerebro de una persona viva, que es psicológicamente continua conmigo, yo sigo existiendo. Esto es cierto con independencia de lo que le ocurra al resto de mi cuerpo. Cuando me trasplantan el corazón de alguien, yo soy el receptor superviviente, no el donante muerto. Cuando se trasplanta mi cerebro en el cuerpo de alguien, puede parecer que yo soy aquí el donante muerto. Pero en realidad aún soy el receptor, y el superviviente. Recibir un nuevo cráneo y un nuevo cuerpo no es más que el caso límite de recibir un nuevo corazón, nuevos pulmones, nuevos brazos, etc. [38].

Desde luego que tendrá su importancia cómo es mi nuevo cuerpo. Si mi nuevo cuerpo fuera totalmente diferente de mi viejo cuerpo, esto afectaría lo que yo podría hacer, con lo que, tal vez, indirectamente, produciría cambios en mi carácter. Pero no hay razón alguna para suponer que ser trasplantado en un cuerpo muy diferente interrumpiría mi continuidad psicológica.

Se ha objetado que «la posesión de algunas clases de rasgos de carácter requiere la posesión de la clase de cuerpo adecuada». Quinton responde a esta objeción. Escribe, de un caso improbable,

> «Sería extraño para una niña de seis años manifestar el carácter de Winston Churchill, sin duda extraño hasta lo monstruoso, pero no

[38] Sigo a Shoemaker (1), p. 22.

totalmente inconcebible. En un principio, evidentemente, que la niña manifestara una resistencia tenaz, una amplitud histórico-mundial de perspectivas, etc., le impresionaría a uno como algo desagradable y pretencioso en una niña tan pequeña. Pero si ella siguiera por este camino, la impresión acabaría por borrarse» [39].

Y lo que es más importante, como argumenta Quinton, esta objeción podría mostrar sólo que tal vez importaría el que mi cerebro fuese alojado en un cierto *tipo* de cuerpo. Pero no podría mostrar que fuese importante que fuera alojado en un cuerpo *concreto*. Y en mi caso imaginario mi cerebro será alojado en un cuerpo que, aunque no sea numéricamente idéntico a mi viejo cuerpo, es muy similar: se trata del cuerpo de mi gemelo.

Según todas las versiones del criterio psicológico, la persona resultante sería yo. Y a la mayor parte de los que creen en el criterio físico se les podría persuadir de que, en este caso, esto es cierto. Como he dicho, el criterio físico debería poner como requisito únicamente la existencia continua de una cantidad *suficiente* de mi cerebro, como para que sea el cerebro de una persona viva, contando con que nadie más tiene suficiente cantidad de este cerebro. Esto haría que fuese yo el que se despertara después de la operación. Y si el cuerpo de mi gemelo fuese exactamente igual al mío, ni siquiera yo podría notar que tengo un cuerpo nuevo.

En efecto, es cierto que es suficiente con un hemisferio. Hay mucha gente que ha sobrevivido, cuando un derrame cerebral o una lesión puso fuera de juego uno de sus hemisferios. Con el hemisferio que le queda, la persona puede que necesite reaprender algunas cosas, como el habla adulta, o la manera de controlar las dos manos. Pero esto es posible. En mi ejemplo estoy asumiendo que, igual que puede ser cierto de algunas personas reales, mis dos hemisferios tienen la gama completa de destrezas. Por eso yo podría sobrevivir con cualquiera de ellos, sin ninguna necesidad de reaprendizaje.

Ahora combinaré estas dos últimas afirmaciones. Yo sobreviviría si mi cerebro fuese trasplantado con éxito al cuerpo de mi gemelo. Y yo podría sobrevivir con sólo la mitad de mi cerebro, habiendo sido destruida la otra mitad. Dados estos dos hechos, está claro que yo sobreviviría si la mitad de mi cerebro fuese trasplantada con éxito al cuerpo de mi gemelo, y la otra mitad fuese destruida.

[39] Quinton, en Perry (1), p. 60.

¿Y qué si la otra mitad *no* fuese destruida? Este es el caso que describió Wiggins: el caso en el que una persona se divide como una ameba [40]. Para simplificarlo, asumo que soy uno de tres trillizos idénticos. Consideremos

Mi División. Mi cuerpo resulta fatalmente herido, como también los cerebros de mis dos hermanos. Me dividen el cerebro, y cada mitad se trasplanta con éxito al cuerpo de uno de mis hermanos. Cada una de las personas resultantes cree que es yo, parece recordar haber vivido mi vida, tiene mi carácter, y es de todas las demás maneras psicológicamente continua conmigo. Y tiene un cuerpo que se parece mucho al mío.

Es probable que este caso siga siendo imposible. Aunque se dice que, en ciertas personas, los dos hemisferios pueden tener la misma gama entera de capacidades, esta afirmación podría ser falsa. Estoy asumiendo aquí que es verdadera cuando se me aplica a mí. También estoy asumiendo que sería posible conectar medio cerebro trasplantado a los nervios de un cuerpo nuevo. Y además que podríamos dividir no sólo los hemisferios superiores, sino también el cerebro inferior. Mis dos primeras asunciones a lo mejor pueden realizarse si hay progreso suficiente en la neurofisiología. Pero parece probable que nunca sería posible dividir el cerebro inferior sin que se viera afectado su funcionamiento.

¿Acaso importa el que, por esta razón, este caso imaginario de una división completa siempre vaya a seguir siendo imposible? Dados los objetivos de mi discusión, no importa. La imposibilidad es meramente técnica. El único rasgo del caso que podría considerarse *radicalmente* imposible —la división de la conciencia de una

[40] Wiggins (1), p. 50. Yo decidí estudiar filosofía casi únicamente porque me quedé cautivado por el caso imaginario de Wiggins.

persona en dos corrientes separadas— es el rasgo que ha ocurrido en la realidad. Habría tenido su importancia que esto hubiese sido imposible, puesto que entonces se podría haber dado apoyo a una afirmación sobre lo que realmente somos. Se podría haber apoyado la afirmación de que somos Egos Cartesianos indivisibles. Por tanto, sí que importa que la división de la conciencia de la persona sea de hecho posible. Parece que no hay una conexión similar entre una tesis determinada sobre lo que somos en realidad y la imposibilidad de dividir y trasplantar con éxito las dos mitades del cerebro inferior. De manera que esta imposibilidad no nos da ninguna razón para negarnos a considerar el caso imaginario en que suponemos que esto sí que puede hacerse. Y considerar este caso nos puede ayudar a decidir tanto lo que creemos ser nosotros mismos como lo que efectivamente somos. Como demostró el ejemplo de Einstein, puede ser útil considerar experimentos mentales imposibles.

Puede que sea útil establecer con antelación lo que creo que este caso demuestra. Proporciona un argumento adicional contra la tesis de que somos entidades que existen separadamente. Pero la conclusión principal que sacaremos es que *la identidad personal no es lo que importa.*

Es natural creer que lo que importa es nuestra identidad. Reconsideremos el caso de la línea secundaria, donde hablé con mi Réplica en Marte, estando yo a punto de morir. Supongamos que creemos que mi Réplica y yo somos dos personas diferentes. Entonces es natural asumir que mi futuro es casi tan malo como la muerte corriente. En unos pocos días, no habrá nadie vivo que sea yo. Es natural asumir que *esto* es lo que importa. Al discutir mi división, comenzaré por hacer esta asunción.

En este caso, cada mitad de mi cerebro será trasplantada con éxito al cuerpo, muy similar, de uno de mis dos hermanos. Las dos personas resultantes serán, desde el punto de vista psicológico, completamente continuas conmigo como yo soy ahora. ¿Qué es lo que me ocurre?

Sólo hay cuatro posibilidades: (1) no sobrevivo; (2) sobrevivo como una de las dos personas; (3) sobrevivo como la otra; (4) sobrevivo como las dos.

Vayamos con la objeción a (1). Yo sobreviviría si mi cerebro fuese trasplantado con éxito. Y la gente ha sobrevivido, de hecho, con la mitad del cerebro destruida. Dados estos hechos, está claro que yo sobreviviría si la mitad de mi cerebro fuese trasplantado con éxito, y la otra mitad fuese destruida. ¿De manera que cómo puedo dejar de sobrevivir si la otra mitad fuese también trasplantada con éxito? ¿Cómo podría un doble éxito ser un fracaso?

Consideremos las dos posibilidades siguientes. Quizás un éxito sea la puntuación máxima. Quizás yo seré una de las dos personas resultantes. Aquí la objeción es que, en este caso, cada mitad de mi cerebro es exactamente igual, y así, para empezar, lo es cada persona resultante. Dados estos hechos, ¿cómo puedo sobrevivir como sólo una de las dos personas? ¿Qué es lo que puede hacerme una de ellas en vez de la otra?

Estas tres posibilidades no pueden descartarse como incoherentes. Podemos entenderlas. Pero, mientras asumamos que la identidad es lo que importa, (1) no es plausible. Mi división no sería tan mala como la muerte. Ni tampoco son (2) y (3) plausibles. Queda la cuarta posibilidad: que yo sobreviva como las dos personas resultantes.

Esta posibilidad podría describirse de varias maneras. Podría afirmar en primer lugar: «Lo que hemos llamado "las dos personas resultantes" no son dos personas. Son una persona. Yo sobrevivo a esta operación. Su efecto es darme dos cuerpos y una mente dividida».

Esta afirmación no puede descartarse absolutamente. Como sostuve antes, debemos admitir la posibilidad de que una persona pudiera tener una mente dividida. Si esto es posible, cada mitad de mi mente dividida podría controlar su propio cuerpo. Pero, aunque esta descripción del caso no pueda rechazarse como inconcebible, conlleva una gran distorsión de nuestro concepto de persona. En mi imaginario Examen de Física mantuve que este caso implicaba sólo a una persona. Había dos rasgos del caso que hacían esto plausible. La mente dividida se reunificaba pronto, y había sólo un cuerpo. Pero si una mente estuviera dividida permanentemente, y sus dos mitades se desarrollaran de formas diferentes, se haría menos plausible sostener que el caso implicaba sólo a una persona. (Recor-

demos al paciente real que se quejaba de que, cuando abrazaba a su mujer, su mano izquierda la apartaba de un empujón.)

El caso de la división completa, donde hay también dos cuerpos, parece haber cruzado el límite con toda claridad. Después de que hube pasado por esta operación, cada uno de los dos «productos» tienen todos los rasgos de una persona. Podrían vivir en los extremos opuestos de la Tierra. Supongamos que tienen pobres recuerdos, y que su apariencia cambia de diferentes formas. Después de muchos años, podrían encontrarse de nuevo, sin ser capaces siquiera de reconocerse. Podríamos tener que afirmar de este par, cuando se pusieran a jugar inocentemente al tenis: «Lo que ves allí es una persona única, jugando al tenis consigo misma. En cada mitad de su mente cree erróneamente que está jugando al tenis con alguien diferente». Si todavía no somos reduccionistas, pensaremos que hay una respuesta verdadera a la pregunta de si estos dos jugadores de tenis son una única persona. Dado lo que queremos decir con «persona», la respuesta tiene que ser No. No puede ser cierto que lo que yo creo que es un desconocido, que está allí detrás de la red, sea en realidad otra parte de mí mismo.

Supongamos que admitimos que los dos «productos» son lo que parecen ser, dos personas diferentes. ¿Podríamos afirmar aún que yo sobrevivo como los dos? Hay otro modo en que podríamos. Yo podría decir: «Sobrevivo a la operación como dos personas diferentes. Pueden ser personas diferentes y sin embargo ser yo, del mismo modo que las tres coronas del Papa forman juntas una corona» [41].

Esta afirmación es también coherente. Pero de nuevo distorsiona enormemente el concepto de persona. Coincidimos satisfechos en que las tres coronas del Papa, cuando se juntan, son una cuarta corona. Pero es difícil de creer que dos personas, juntas, sean una tercera persona. Supongamos que las personas resultantes se enfrentaran en un duelo. ¿Hay tres personas luchando, una de cada lado y una tercera en ambos? Y supongamos que una de las balas

[41] Cf. Wiggins (1), p. 40. Debo esta sugerida manera de hablar, y una de las objeciones a ella, a Michael Woods.

mata «a alguien». ¿Hay dos actos, un asesinato y un suicidio? ¿Cuántas personas quedan vivas? ¿Una o dos? La tercera persona compuesta no tiene vida mental separada. Es difícil de creer que habría realmente tal tercera persona. En vez de decir que las personas resultantes juntas me constituyen —de forma que el par es un trío— es mejor tratarlas como un par y describir la relación que tienen conmigo de un modo más simple.

Podrían decirse otras cosas. Podría sugerirse que las dos personas resultantes son *ahora* diferentes personas, pero que, antes de mi división, *eran* la misma persona. Antes de mi división, eran yo. Esta sugerencia es ambigua. La afirmación puede ser que, antes de mi división, las dos *juntas* eran yo. Según esta explicación, había tres personas diferentes aun antes de mi división. Esto es incluso menos plausible que la afirmación que acabo de rechazar. (Podría pensarse que he entendido mal esta sugerencia. La afirmación puede ser que las personas resultantes no existiesen, como personas separadas, antes de mi división. Pero si entonces no existían, no puede haber sido verdadero que las dos juntas fueran yo.)

Puede en cambio sugerirse que, antes de mi división, *cada una* de las personas resultantes *era* yo. Después de mi división, ninguna es yo, puesto que ahora yo no existo. Pero, si cada una de esas personas *era* yo, cualquier cosa que me ocurriera a mí le tiene que haber ocurrido a cada una de esas personas. Si yo no sobreviví a mi división, ninguna de ellas sobrevivió. Como *hay* dos personas resultantes, el caso implica a *cinco* personas. Esta conclusión es absurda. ¿Podemos negar la asunción que implica esta conclusión? ¿Podemos mantener que, aunque cada persona resultante *era* yo, lo que me ocurrió a mí no le ocurrió a esas personas? Asumamos que todavía no me he dividido. Según esta sugerencia, es ahora cierto que cada una de las personas resultantes *es* yo. Si lo que me ocurre a mí no le ocurre a X, X no puede ser yo.

Hay formas descabelladas de negar esta última afirmación. Formas que apelan a afirmaciones sobre la identidad temporalizada. Llamemos a una de las personas resultantes *Izquierdo*. Yo podría preguntar, «¿Son *Izquierdo* y *Derek Parfit* nombres de la misma persona?». Para los que creen en la identidad temporalizada, esta no es una pre-

gunta adecuada. Como esto muestra, las afirmaciones sobre identidad temporalizada son radicalmente diferentes del modo en que pensamos ahora. Me limitaré a declarar aquí lo que creo que otros han demostrado: que estas afirmaciones no resuelven nuestro problema.

David Lewis hace una propuesta diferente. Según su idea, hay dos personas que comparten mi cuerpo incluso antes de mi división. En sus detalles, esta propuesta es a la vez elegante e ingeniosa. No repetiré aquí por qué, como he dicho en otra parte, tampoco resuelve nuestro problema [42].

He discutido varias opiniones nada corrientes sobre lo que ocurre cuando me divido. Según ellas, el caso implica a una única persona, un dúo, un trío dos de cuyos integrantes forman el tercero, y un quinteto. Sin duda, podríamos hacer aparecer al cuarteto que falta. Pero sería tedioso considerar más opiniones de estas. Todas conllevan distorsiones demasiado grandes del concepto de persona. Debemos, por tanto, rechazar la cuarta posibilidad sugerida: la afirmación de que, de alguna manera, yo sobrevivo como las dos personas resultantes.

Hay otras tres posibilidades: que yo seré *una*, o *la otra* o *ninguna* de esas personas. Estas tres afirmaciones parecen inverosímiles. Nótese además que, como antes, no podríamos *averiguar* lo que ocurre aunque pudiéramos llevar a cabo realmente la operación. Supongamos, por ejemplo, que sobrevivo como una de las personas resultantes. Yo pensaría que he sobrevivido. Pero sabría que la otra persona resultante cree equivocadamente que es yo, y que ha sobrevivido. Como yo sabría esto, no podría confiar en mi propia creencia. Yo podría ser la persona resultante con la falsa creencia. Y como los dos afirmaríamos ser yo, los demás no tendrían ninguna razón para creer a uno en vez de creer al otro. Aunque lleváramos a cabo esta operación, con ello no aprenderíamos nada.

[42] Véanse «Survival and Identity» [«Supervivencia e identidad»], de Lewis, y mi «Lewis, Perry, and What Matters» [«Lewis, Perry y lo que importa»], los dos en Rorty.

Fuese lo que fuese lo que me ocurrió a mí, no podríamos descubrir qué ocurrió. Esto sugiere una respuesta más radical a nuestra pregunta. Sugiere que la concepción reduccionista es verdadera. Quizás no haya aquí posibilidades diferentes, cada una de las cuales podría ser lo que ocurre, aunque nosotros nunca pudiésemos saber cuál se da en la realidad. Quizás, cuando sabemos que cada persona resultante tendría la mitad de mi cerebro, y sería psicológicamente continua conmigo, lo sabemos todo. ¿Qué estamos suponiendo cuando sugerimos, por ejemplo, que una de las personas resultantes podría ser yo? ¿Qué haría a esta la respuesta verdadera?

Pienso que no puede haber diferentes posibilidades, tales que cada una de ellas podría ser la verdad, a no ser que seamos entidades que existen separadamente, como Egos Cartesianos. Si lo que realmente soy es un Ego particular, esto explica cómo puede ser cierto que una de las personas resultantes sea yo. Puede ser cierto que sea en el cerebro y en el cuerpo de esta persona donde este Ego particular recobró la conciencia.

Si creemos en Egos Cartesianos, nos podríamos acordar del asno de Buridán, que murió de hambre entre dos fardos de heno igualmente nutritivos. El asno no tenía ninguna razón para comer de uno de los fardos y no del otro. Siendo una bestia hiper-racional, rehusó hacer una elección para la que no había razón suficiente. En mi ejemplo, no habría ninguna razón por la que el Ego particular que soy debiera despertar como una de las dos personas resultantes, sino que esto podría ocurrir sólo de un modo azaroso, como se dice en el caso de las partículas fundamentales.

La pregunta más difícil para los que creen en Egos Cartesianos es si yo sobreviviría en absoluto. Como cada una de las personas resultantes sería psicológicamente continua conmigo, no habría evidencia que apoyara ninguna respuesta a la pregunta. Este argumento conserva su fuerza aunque yo sea un Ego Cartesiano.

Como antes, un cartesiano podría objetar que he descrito mal lo que ocurriría. Podría decir que, si lleváramos a la práctica la operación, no sería cierto efectivamente que *las dos* personas resultantes fuesen psicológicamente continuas conmigo. Podría ser cierto que una u otra de estas personas fuera psicológicamente continua con-

migo. En cualquiera de los casos, la persona sería yo. Podría en cambio ser cierto que ninguna persona fuera psicológicamente continua conmigo. En este caso, yo no sobreviviría. En cada uno de estos tres casos, aprenderíamos la verdad.

Que esta sea una buena objeción depende de cuál es la relación entre nuestros rasgos psicológicos y los estados de nuestros cerebros. Como he dicho, tenemos evidencia concluyente de que el portador de la continuidad psicológica *no* es indivisible. En los casos reales en que se han desconectado los hemisferios se produjeron como consecuencia dos series de pensamientos y sensaciones. Estas dos corrientes de conciencia eran ambas psicológicamente continuas con la corriente original. La continuidad psicológica ha adoptado por tanto, en varios casos reales, una forma dividida. Este hecho refuta la objeción que se acaba de dar, y justifica mi afirmación de que, en el caso imaginario de mi división, las dos personas resultantes serían psicológicamente continuas conmigo. Como esto es así, la Concepción Cartesiana puede plantearse aquí sólo en la versión más dudosa que no conecta el Ego con hecho ninguno, ni observable ni introspeccionable. Aunque yo fuera un Ego semejante, nunca podría saber si había o no había sobrevivido. Para los cartesianos, este caso es un problema sin solución posible.

Supongamos que, por las razones dadas anteriormente, rechazamos la idea de que cada uno de nosotros es realmente un Ego Cartesiano. Y rechazamos la idea de que una persona es cualquier otra clase de entidad que exista separadamente, aparte de su cerebro y su cuerpo, y de varios sucesos mentales y físicos. ¿Cómo deberíamos contestar entonces a la pregunta por lo que sucede cuando me divido? Distinguí cuatro posibilidades. Cuando discutí cada posibilidad, parecieron presentarse fuertes objeciones a la afirmación de que esto sería lo que ocurre. Si pensamos que son posibilidades diferentes, cualquiera de las cuales podría ser lo que ocurre, el caso es un problema también para nosotros.

Si hacemos nuestra la Concepción Reduccionista, el problema desaparece. Para ella, las afirmaciones que he discutido no describen

posibilidades diferentes, de las cuales cualquiera podría ser verdadera, y una tiene que ser verdadera. Las afirmaciones nada más que son descripciones diferentes del mismo resultado. Sabemos cuál es este resultado. Habrá dos personas futuras, cada una de las cuales tendrá el cuerpo de uno de mis hermanos y será completamente continua conmigo desde el punto de vista psicológico, porque tiene la mitad de mi cerebro. Sabiendo esto, lo sabemos todo. Es verdad que puedo preguntar, «Pero ¿seré una de estas dos personas, o la otra, o ninguna de las dos?». Pero debería considerar esta pregunta como una pregunta vacía. Aquí tenemos una pregunta similar: En 1881 el Partido Socialista Francés se escindió. ¿Qué ocurrió? ¿El Partido Socialista Francés dejó de existir, o siguió existiendo como uno u otro de los dos nuevos Partidos? Dados ciertos detalles adicionales, esta sería una pregunta vacía. Aunque no tengamos una respuesta para esta pregunta, podríamos saber exactamente lo que ocurrió.

Ahora tengo que distinguir dos maneras en que una pregunta puede ser vacía. En relación con algunas preguntas debemos afirmar que son vacías y además que no tienen respuesta. Podríamos decidirnos a *dar* respuestas a estas preguntas. Pero podría ser cierto que cualquier respuesta posible sería arbitraria. Si es así, no vendría a cuento para nada y hasta podría ser engañoso dar una respuesta semejante. Esto ocurriría con la pregunta «¿Sobreviviré?» en los casos centrales del espectro combinado. Y ocurriría en los casos centrales de los otros espectros, si yo no sobreviviera en el caso del extremo lejano.

Y hay otra clase de casos en que una pregunta puede ser vacía. En estos casos hay un sentido en que la pregunta tiene una respuesta. Es vacía porque no describe diferentes posibilidades, de las que cualquiera podría ser verdadera y una tiene que serlo. La pregunta se limita a darnos descripciones diferentes del mismo resultado. Podríamos saber toda la verdad acerca de este resultado sin elegir una de estas descripciones. Pero si decidimos dar una respuesta a esta pregunta vacía, lo hacemos porque hay una de estas descripciones que es mejor que las demás. Como es así, podemos decir que esta descripción es la respuesta a esta pregunta. Y yo afir-

mo que hay una descripción del caso en que me divido que es la mejor. La mejor de las descripciones es que ninguna de las personas resultantes será yo.

Como este caso no implica diferentes posibilidades, la pregunta importante no es, «¿Cuál es la mejor descripción?». La pregunta importante es : «¿Qué debe importarme? ¿Cómo debo considerar la perspectiva de la división? ¿La debería considerar como la muerte, o como la supervivencia?». Una vez que hayamos contestado a esta pregunta, podremos decidir si he dado la mejor descripción de todas.

Antes de discutir lo que importa, cumpliré una promesa anterior. Una objeción contra el criterio psicológico es que la continuidad psicológica presupone la identidad personal. Respondí a esta objeción, en el caso de la memoria, apelando al concepto más amplio de cuasi-memoria. Jane cuasi-recordaba tener las experiencias pasadas de alguien distinto. Mi división proporciona otro ejemplo. Puesto que una de las dos personas resultantes, como mínimo, no será yo, podrá cuasi-recordar haber vivido la vida de alguien distinto.

No demostré que, al describir las otras relaciones que están implicadas en la continuidad psicológica, no necesitáramos presuponer identidad personal. Ahora que he descrito mi división, esto puede demostrarse fácilmente. Una relación directa distinta es la que se da entre una intención y la acción posterior en que esta intención se realiza. Puede que sea una verdad lógica el que sólo podemos intentar llevar a cabo nuestras propias acciones. Pero podemos usar un nuevo concepto de *cuasi-intención*. Una persona podría cuasi-intentar realizar las acciones de otra persona. Cuando esta relación se da, no presupone identidad personal.

El caso de la división muestra lo que esto implica. Yo podría cuasi-intentar que una de las personas resultantes vagara por el mundo y también que la otra se quedara en casa. Lo que yo cuasi-intento será llevado a cabo, no por mí, sino por las dos personas resultantes. Normalmente, si intento que alguien distinto haga algo, no puedo conseguir que lo haga simplemente formándome esta intención. Pero si estoy a punto de dividirme, bastaría simplemente con formar cuasi-intenciones. Las dos personas resultantes heredarían estas cuasi-intenciones, y, a no ser que cambiaran su opinión heredada, las llevarían a cabo. Como podrían cambiar de opinión, yo no podría tener la seguridad de que harían lo que cuasi-intenté. Pero ocurre lo mismo en mi propia vida. Como puedo cambiar de opinión, no puedo tener la seguridad de que haré lo que ahora tengo la intención de hacer. Pero tengo cierta capacidad de controlar mi futuro formándome intenciones firmes. Si estuviera a punto de dividirme, tendría esa misma capacidad, por la formaciones de cuasi-intenciones, para controlar los futuros de las dos personas resultantes.

Parecidas observaciones se aplican a todas las demás conexiones psicológicas directas, como las que están implicadas en la continuidad de carácter. Todas estas conexiones se dan entre yo y cada una de las personas resultantes. Como al menos una de estas personas no puede ser yo, ninguna de estas conexiones presupone identidad personal.

90. ¿QUÉ ES LO QUE IMPORTA CUANDO ME DIVIDO?

Hay quienes considerarían la división tan mala o casi tan mala como la muerte corriente. Esta reacción es irracional. Debemos considerar la división casi tan buena como la supervivencia corriente. Como he sostenido, los dos «productos» de esta operación serían dos personas diferentes. Consideremos mi relación con cada una de estas personas. ¿Esta relación deja de contener algún elemento vital que esté contenido en la supervivencia corriente? Está claro que no. Yo sobreviviría si estuviera en esta misma relación con sólo una de las personas resultantes. Es un hecho que uno puede sobrevivir aunque la mitad de su cerebro sea destruida. Y, después de reflexionar, quedó claro que yo sobreviviría si mi cerebro entero fuera trasplantado con éxito en el cuerpo de mi hermano. Quedó claro, por consiguiente, que yo sobreviviría si la mitad de mi cerebro fuese destruida y la otra mitad trasplantada con éxito en el cuerpo de mi hermano. En el caso que estamos considerando ahora, mi relación

con cada una de las personas resultantes contiene, entonces, todo lo que yo necesitaría para sobrevivir como esa persona. No puede ser la *naturaleza* de mi relación con cada una de las personas resultantes la que, en este caso, cause que deje de haber supervivencia. No *falta* nada. Lo que falla sólo puede ser la duplicación.

Supongamos que acepto esto, pero todavía considero la división casi tan mala como la muerte. Mi reacción es ahora injustificable. Soy como alguien que, cuando se le dice que una droga podría doblar sus años de vida, considera que tomar esa droga es como la muerte. La única diferencia en el caso de la división es que los años extra van a transcurrir simultáneamente. Esta es una diferencia interesante; pero no puede significar que *no* haya años que van a transcurrir. Podríamos decir por nuestra parte: «Vas a perder tu identidad. Pero hay diferentes maneras en que esto puede ocurrir. Morir es una, dividirse es otra. Considerarlas lo mismo es confundir dos con cero. La supervivencia doble no es lo mismo que la supervivencia corriente. Pero eso no la iguala a la muerte. La hace algo todavía más diferente de la muerte».

El problema con la doble supervivencia es que no se ajusta a la lógica de la identidad. Como otros reduccionistas, afirmo que

> *La relación R* es lo que importa. R es conexividad psicológica y/o continuidad psicológica, con la clase correcta de causa [43].

También afirmo que

> En una explicación de lo que importa, la clase correcta de causa podría ser cualquier causa.

Otros reduccionistas podrían poner el requisito de que R tenga una causa fiable, o tenga su causa normal. Para posponer este desa-

[43] Otros Reduccionistas con quienes en conjunto coincido serían H. P. Grice [en Perry (1)], A. J. Ayer [véase especialmente «The Concept of a Person» («El concepto de persona»)], en Ayer (1), A. Quinton, J. L. Mackie, [en Mackie (4) y (5)], J. Perry, especialmente en «The Importance of Being Identical» [«La importancia de ser idéntico»], en Rorty, y en Perry (29, D. K. Lewis (en Rorty), y S. Shoemaker [en su *Personal Identity* (*La identidad personal*), Blackwell, 1984].

cuerdo, consideraré sólo casos en que R tenga su causa normal. En estos casos, todos los reduccionistas aceptarían la siguiente afirmación: una persona futura será yo si va a estar R-relacionada conmigo como soy ahora, y ninguna persona diferente va a estar R-relacionada conmigo. Si no hay tal persona diferente, el hecho de que esta persona futura va a ser yo sólo consiste en el hecho de que la relación R se da entre nosotros. No hay nada más en la identidad personal que el darse de la relación R. En casi todos los casos reales, R toma la forma uno-uno. Se da entre una persona que existe en el presente y una persona futura. Cuando R toma la forma uno-uno, podemos usar el lenguaje de la identidad. Podemos decir que esta persona futura será esta persona presente.

En el caso imaginario en que me divido, R toma una forma ramificada. Pero la identidad personal no puede tomar una forma ramificada. Yo y las dos personas resultantes no podemos ser la misma persona. Como no puedo ser idéntico a dos personas diferentes, y sería arbitrario llamar yo a una de ellas, lo mejor que podemos hacer es describir el caso diciendo que ninguna va a ser yo.

¿Cuál es la relación que es importante? ¿Lo que importa es la identidad personal o la relación R? En los casos corrientes no tenemos que decidir cuál de ellas es la que importa puesto que las dos relaciones coinciden. En el caso de mi división las dos relaciones no coinciden. Por tanto tenemos que decidir cuál de las dos es la que importa.

Si creemos ser entidades que existen separadamente, podríamos afirmar de manera verosímil que la identidad es lo que importa. Según esta forma de ver las cosas, la identidad personal es un hecho adicional profundo. Pero tenemos evidencia suficiente para rechazar esta concepción. Si somos reduccionistas, no podemos afirmar de manera convincente que, de las dos relaciones, es la identidad lo que importa. Según nuestra idea, el hecho de la identidad personal no consiste más que en el darse de la relación R, cuando toma una forma no ramificada. Si la identidad personal sólo consiste en esta otra relación, esta otra relación tiene que ser lo que importa.

Puede objetarse: «Estás equivocado al afirmar que no hay nada más en la identidad que la relación R. Como tú mismo has dicho,

la identidad personal tiene un rasgo extra, que no está contenido en la relación R. La identidad personal consiste en que R se da *de forma única* —que se da entre una persona presente y *sólo una* persona futura—. Como hay algo más en la identidad personal que en la relación R, podemos afirmar racionalmente que, de las dos, es la identidad lo que importa».

Al responder a esta objeción, será útil usar algunas abreviaturas. Llamemos a la identidad personal *IP*. Cuando una relación se da de forma única, o en la forma uno-uno, llamaremos a este hecho *U*. La tesis que acepto puede formularse con esta fórmula:

$$IP = R + U$$

La mayoría de nosotros estamos convencidos de que IP importa, o tiene valor. Asumamos que R puede tener también valor. Entonces hay cuatro posibilidades:

(1) R sin U no tiene valor.
(2) U aumenta el valor de R, pero R tiene valor incluso sin U.
(3) U no supone ninguna diferencia para el valor de R.
(4) U reduce el valor de R (pero no lo bastante como para eliminarlo, puesto que R + U = IP, que tiene valor).

¿La presencia o ausencia de U puede significar una gran diferencia para el valor de R? Como defenderé, esto no es verosímil. Si voy a estar R-relacionado con una persona futura, la presencia o ausencia de U no significa diferencia alguna para la naturaleza intrínseca de mi relación con esta persona. Y lo que importa por encima de todo tiene que ser la naturaleza intrínseca de esta relación.

Como esto es así, R sin U todavía tendría al menos la mayor parte de su valor. Sumar U hace a R = IP. Si sumar U no incrementa mucho el valor de R, R tiene que ser lo que importa fundamentalmente, e IP importa sobre todo por la presencia de R. Si U no representa ninguna diferencia para el valor de R, IP importa sólo a causa de la presencia de R. Como U puede decirse convincentemente que introduce una pequeña diferencia, IP puede que tenga,

comparada con R, algún valor extra. Pero este valor sería mucho menor que el valor intrínseco de R. El valor de IP es mucho menor que el valor que R tendría en ausencia de IP, cuando U deja de darse.

Si fuera propuesta por sí misma, sería difícil aceptar la idea de que la identidad personal no es lo que importa. Pero creo que, cuando consideramos el caso de la división, esta dificultad desaparece. Cuando vemos *por qué* ninguna persona resultante será yo, pienso que, tras reflexionar, podemos ver también que esto no importa, o importa sólo un poco.

El caso de la división apoya parte de la Concepción Reduccionista: la afirmación de que nuestra identidad no es lo que importa. Pero no apoya otra tesis reduccionista: que nuestra identidad pueda ser indeterminada. Si abandonamos la idea de que la identidad es lo que importa, podemos afirmar que *hay* aquí una respuesta a mi pregunta. Ninguna de las personas resultantes será yo. Estoy a punto de morir. Mientras creíamos que la identidad es lo que importa, esta afirmación tenía la implicación, inverosímil, de que yo debo considerar mi división sin duda tan mala como la muerte corriente. Pero lo inverosímil desaparece si afirmamos en cambio que este modo de morir es casi tan bueno como la supervivencia corriente.

Hay todavía espacio para desacuerdos menores. Aunque R es lo que fundamentalmente importa, U puede representar una leve diferencia. Yo podría considerar mi división de algún modo mejor que la supervivencia ordinaria, o de algún modo peor.

¿Por qué podría yo pensar que es de algún modo peor? Podría afirmar que la relación entre yo y cada una de las personas resultantes no es en absoluto la relación que importa en la supervivencia corriente. Y no porque falte algo, sino porque la división nos aporta *demasiado*. Puedo pensar que cada persona resultante, en un respecto, tendrá una vida que es peor que la mía. Cada una tendrá que vivir en un mundo en que hay alguien más que, al menos para empezar, es exactamente como ella misma. Esto tal vez resulte desagradablemente asombroso. Y hará surgir problemas prácticos. Supongamos que mi mayor deseo es escribir determinado libro.

Esto sería lo que cada persona resultante querría hacer en mayor medida. Pero carecería de sentido que las dos escribiéramos ese libro. Sería absurdo que las dos hiciéramos lo que tenemos más ganas de hacer.

Consideremos a continuación las relaciones entre las personas resultantes y la mujer que amo. Puedo asumir que, como ella me ama, las amará a las dos. Pero no podría dedicar a los dos la atención indivisa que nos dedicamos ahora cada uno al otro.

De estas y de otras maneras las vidas de las personas resultantes puede que no fueran tan buenas como la mía. Lo cual podría justificar que yo considerara mi división no tan buena como la supervivencia corriente. Pero no podría justificar el considerar mi división mucho menos buena, o tan mala como la muerte. Y debemos notar que este razonamiento ignora el hecho de que estas dos vidas, tomadas juntas, serían el doble de largas que el resto de la mía.

En lugar de considerar la división como de algún modo peor que la supervivencia corriente, podría considerar que es mejor. La razón más simple sería la que acabamos de dar: se doblarían los años que quedan por vivirse. Podría tener razones más particulares. Por ejemplo, podría haber dos carreras de la longitud de una vida, que yo deseara ardientemente seguir. Podría desear ardientemente ser novelista y filósofo. Si me divido, cada persona resultante podría seguir una de estas carreras. Y cada una se alegraría de que la otra tuviera éxito. Del mismo modo que nos pueden dar orgullo y alegría los éxitos de nuestros hijos, cada persona resultante estaría orgullosa y contenta de los éxitos de la otra.

Si tengo dos ambiciones poderosas pero incompatibles, la división me proporciona un modo de realizarlas, de una manera que contentaría a cada persona resultante. Esta es una forma en que la división podría ser mejor que la supervivencia corriente. Pero hay otros problemas que la división no podría resolver del todo. Supongamos que me debato entre una obligación desagradable y un deseo seductor. Yo no podría resolver el problema completamente con la cuasi-intención de que una de las personas resultantes cumpla con mi obligación y con la cuasi-intención de que la otra haga lo que deseo. La persona resultante a quien dirijo mi cuasi-intención de cumplir con mi obligación se debatiría ella misma entre la obligación y el deseo. ¿Por qué debería ser *ella* la que cumpliese mi desagradable obligación? Podemos prever que aquí surgirá el problema. Mi obligación podría cumplirse si el deseo seductor no pudiera ser realizado por más de una persona. Podría ser el deseo de fugarse con alguien que quiere sólo a uno de compañero. El que fracase en esta competición podría entonces, a regañadientes, cumplir con mi obligación. Mi problema se resolvería, aunque de una manera menos atractiva.

Estas observaciones les parecerán absurdas a los que todavía no han sido convencidos de que la Concepción Reduccionista es verdadera, o de que la identidad no es lo que importa. Una de estas personas tal vez diría: «Si yo no voy a *ser* ninguna de las personas resultantes, la división no podría realizar mis ambiciones. Aunque una de las personas resultantes sea un novelista célebre, y la otra un célebre filósofo, esto no realiza ninguna de mis ambiciones. Si una de mis ambiciones es ser un novelista célebre, mi ambición es que *yo* sea un novelista célebre. Ambición que no se realizará si yo dejo de existir y *alguien diferente* es un novelista célebre. Y esto es lo que ocurriría si yo no voy a ser ninguna de las personas resultantes».

Esta objeción da por sentado que hay una cuestión real en si yo seré una de las personas resultantes, la otra o ninguna. Es natural dar por sentado que estas son tres posibilidades diferentes, de forma que cualquiera de las tres podría ser lo que ocurre. Pero, como he defendido, a no ser que yo sea una entidad que existe separadamente, del tipo de un Ego Cartesiano, estas no pueden ser tres posibilidades diferentes. No hay nada que pudiera verificar que cualquiera de las tres podría ser lo que realmente ocurre. (Esto es compatible con mi afirmación de que hay una descripción mejor de este caso: que yo no seré ninguna de las personas resultantes. Esto no me compromete con la idea de que hay posibilidades diferentes: sería así sólo si una de las otras descripciones *pudiera* haber sido la verdad —cosa que niego.)

Podríamos dar una descripción diferente. Podríamos decir que yo seré la persona resultante que se convierte en un novelista célebre.

Pero sería un error pensar que mi ambición se realizaría si y sólo si nosotros *llamáramos* a esta persona resultante yo. Cómo elijamos describir este caso no tiene significación racional ni moral.

Ahora revisaré lo que he afirmado. Cuando discutí los Espectros Psicológico, Físico y Combinado, defendí que nuestra identidad puede ser indeterminada. Esta no es la concepción natural. Estamos inclinados a pensar que, para la pregunta «¿Estoy a punto de morir?», tiene que haber siempre una respuesta que tiene que ser Sí o No. Estamos inclinados a creer que nuestra identidad tiene que ser determinada. Sostuve que esto no puede ser verdadero a no ser que seamos entidades que existen separadamente, del tipo de los Egos Cartesianos. No podemos negar que una persona sea una entidad tal, e insistir además en que la existencia continua de una persona tiene los mismos rasgos especiales que los cartesianos atribuyen a la existencia del Ego. Concedí por mi parte que podríamos haber sido entidades semejantes. Pero hay mucha evidencia contra esta concepción.

Si negamos que seamos entidades que existen separadamente, tenemos, como he afirmado, que hacernos reduccionistas. Una afirmación reduccionista es que podemos imaginar casos en que la pregunta «¿Estoy a punto de morir?» no tiene respuesta, es vacía. Esta parece la idea menos inverosímil en relación con los casos centrales del Espectro Combinado. Otra afirmación reduccionista es que la identidad personal no es lo que importa. Esta parece la idea menos inverosímil en relación con el caso de Mi División. De estas dos afirmaciones reduccionistas, la segunda es la más importante, desde el momento en que se aplica a nuestras propias vidas.

Si aceptamos la Concepción Reduccionista, tendremos más preguntas que responder acerca de lo que importa. Se trata de preguntas acerca de la significación racional y moral de ciertos hechos. Pero, según la Concepción Reduccionista, los así llamados «casos problema» dejan de generar problemas sobre lo que ocurre. Aunque no tengamos respuesta a una pregunta sobre la identidad personal, podemos saberlo todo sobre lo que ocurre.

¿He pasado por alto alguna concepción? He afirmado que, si rechazamos la idea de que somos entidades que existen separadamente, deberíamos aceptar alguna versión de la Concepción Reduccionista. Pero hay autores que defienden modos de pensar que obviamente no son versiones de ninguna de estas concepciones. Por eso discutiré lo que dicen esos autores.

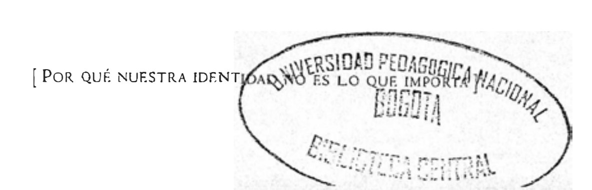
91. POR QUÉ NO HAY CRITERIO DE IDENTIDAD QUE PUEDA CUMPLIR DOS REQUISITOS PLAUSIBLES

Además del discutido en la Sección 83, Williams presenta otro argumento en contra del criterio psicológico. De nuevo puede ser de utilidad que formule de antemano lo que pienso que demuestra este argumento. Williams afirma que el criterio de identidad personal tiene que cumplir dos requisitos. Yo afirmaré que *ningún* criterio de identidad plausible puede cumplirlos los dos. En contraste, según la Concepción Reduccionista, los requisitos análogos pueden cumplirse. Por tanto, el argumento nos da más razones para aceptarla. Al discutir el argumento, dejaré de lado por ello brevemente esta concepción. Puede esperar entre bastidores, para reaparecer cuando la acción lo exija.

El argumento de Williams desarrolla una observación de Reid contra la afirmación de Locke de que quienquiera que «tenga la conciencia de acciones presentes y pasadas es la misma persona a quien pertenecen». Esto implica, escribe Reid, «que si la misma conciencia puede transferirse de un ser inteligente a otro... entonces dos o veinte seres inteligentes pueden ser la misma persona» [44].

Williams argumenta como sigue. La identidad es lógicamente una relación uno-uno. Es lógicamente imposible para una persona ser idéntica a más de una persona. No puedo ser la misma persona que dos personas diferentes. Como hemos visto, la continuidad psicológica no es lógicamente una relación uno-uno. Dos personas futuras diferentes podrían ser psicológicamente continuas conmigo. Como estas personas diferentes no pueden las dos ser yo, la continuidad psicológica no puede ser el criterio de identidad.

[44] Reid, en Perry (1), p. 114.



Williams entonces afirma que, para ser aceptable, un criterio de identidad tiene que ser él mismo lógicamente una relación uno-uno. Tiene que ser una relación que no pueda *de ningún modo* darse entre una persona y dos personas futuras. Por eso afirma que el criterio de identidad no puede ser la continuidad psicológica [45].

Algunos reponen que este criterio podría apelar a la continuidad psicológica *no ramificada*. Esta es la versión del criterio que he discutido. Según lo que llamo el criterio psicológico, una persona futura será yo si va a estar R-relacionada conmigo, y no hay ninguna otra persona que vaya a estar R-relacionada conmigo. Como esta versión del criterio es lógicamente una relación uno-uno, se ha dicho que responde a la objeción de Williams [46].

Williams rechaza esta respuesta. Afirma

El Requisito (1): El que una persona futura sea yo tiene que depender sólo de rasgos *intrínsecos* de la relación entre nosotros. No puede depender de lo que le ocurre a *otras* personas.
El Requisito (2): Como la identidad personal tiene una gran significación, el que la identidad se dé no puede depender de un hecho trivial [47].

Estos requisitos son ambos plausibles. Y la continuidad psicológica no ramificada no cumple ninguno. Por eso rechaza Williams esta versión del Criterio Psicológico.

Esta objeción puede parecer demasiado abstracta para ser convincente. Puede mostrarse su fuerza si varío la historia imaginaria con que empecé. Consideremos el teletransporte simple, donde el escáner destruye mi cerebro y mi cuerpo. Una vez que mi cianotipo es transmitido a Marte, el replicador hace una reproducción orgánica perfecta. Mi Réplica en Marte pensará que es yo, y será de todas las maneras psicológicamente continua conmigo.

Supongamos que aceptamos el Criterio Psicológico que apela a la relación R cuando se da en la forma uno-uno. Y supongamos que aceptamos la versión amplia, que deja que R tenga cualquier causa fiable. Este criterio implica que mi Réplica en Marte será yo. Pero podríamos enterarnos de que mi cianotipo también se está transmitiendo a Io, uno de los satélites de Júpiter. Entonces tenemos que afirmar que seré yo el que se despierte en Marte, y que seguiré existiendo si los científicos de Io ignoran mi cianotipo. Pero como hagan después otra Réplica mía, cuando esa Réplica se despierte yo dejaré de existir. Aunque las personas que me rodean en Marte no notarán ningún cambio, en ese momento vendrá a la existencia en mi cerebro y en mi cuerpo una persona nueva. Williams objetaría que, si *soy yo* el que despierta en Marte, que siga existiendo allí no puede depender, como decimos, de lo que le ocurre a alguien a millones de millas de distancia, cerca de Júpiter. Nuestra afirmación viola el Requisito (1).

Como he defendido, lo que fundamentalmente importa es si yo estaré R-relacionado con al menos una persona futura. Es relativamente trivial si estaré también R-relacionado con alguna otra persona. Según esta versión del Criterio Psicológico, el que yo sea idéntico a una persona futura depende de este hecho relativamente trivial. Esto viola el requisito (2).

Williams añadiría estas observaciones. Una vez que vemos que el teletransporte podría producir muchas Réplicas mías, que serían diferentes personas unas de otras, deberíamos negar que yo fuera a despertar efectivamente en Marte, aunque hicieran sólo una única Réplica. Si hicieran dos Réplicas, las dos no podrían ser yo. Si las dos no podrían ser yo, pero están producidas exactamente de la misma manera, debemos concluir que ninguna sería yo. Pero mi relación con una de las Réplicas es intrínsecamente la misma tanto si hacen la otra como si no. Como la identidad tiene que depender de los rasgos intrínsecos de una relación, yo no sería ninguna Réplica aunque no hicieran la otra [48].

[45] En Williams, reeditado en Williams (2), pp. 19-25.
[46] Shorter; y J. M. R. Jack (inédito), que exige que este criterio se incruste en una teoría causal.
[47] Williams (2), p. 20.

[48] Wiggins (1), (2) y (3) presenta argumentos parecidos. Algunos de los temas planteados, que yo no discuto aquí, se discuten sucintamente en Nozick (3), pp. 656-9.

Williams acude a este argumento para dar apoyo a una versión no reduccionista del Criterio Físico. (Esta versión es no reduccionista porque da por sentado que la identidad personal es un hecho adicional que, más que consistir en continuidad física, la requiere.) Como Williams admite, un argumento similar puede desafiar esta concepción. Él rechaza el Criterio Psicológico porque apela a una relación que puede adoptar una forma ramificada, dándose entre una persona y dos o más personas futuras, y por ello este criterio es incapaz de cumplir sus dos requisitos. Entonces considera la objeción de que su versión del Criterio Físico también es incapaz de cumplirlos. Como él mismo escribe, «Es posible imaginar a un hombre dividiéndose, como una ameba, en dos simulacros de sí mismo» [49].

Williams da dos respuestas a esta objeción. Supongamos que creemos que mi cerebro y mi cuerpo son físicamente continuos con el cerebro y el cuerpo de la persona a quien mis padres cuidaron como a su segundo hijo. Deseamos saber si esta continuidad física asumió una forma anormal, ramificada. Si conociéramos toda la historia de este cerebro y este cuerpo físicamente continuos, esto «revelaría inevitablemente» si se había dado un caso semejante de división como la de la ameba. La afirmación comparable no es verdadera en el caso de la continuidad psicológica. Podríamos conocer toda la historia de la continuidad psicológica entre yo en la Tierra y mi Réplica en Marte, y con todo ser incapaces de saber que yo tengo otra Réplica en Io. La ramificación es un problema para los dos Criterios, el Físico y el Psicológico. Pero este problema es menos serio para el Criterio Físico, puesto que sería en principio más fácil saber cuándo surge.

Williams también afirma que, cuando un objeto físico se divide, esto es un rasgo intrínseco de su continuidad espacio-temporal. En contraste, cuando dos personas son psicológicamente continuas con una persona anterior, este hecho no es un rasgo intrínseco de ninguna de estas relaciones. A diferencia del Criterio Psicológico, el Físico cumple el requisito (1).

[49] Williams (2), p. 23.

Mi división imaginaria proporciona objeciones contra el Criterio Físico. Revisé este criterio de dos modos. Primero consideré el caso en que se trasplanta mi cerebro al cuerpo de mi gemelo idéntico. Después de reflexionar, quedó claro que yo soy aquí el receptor superviviente y no el donante muerto. Si a mi cerebro se le da un cuerpo nuevo, esto no es sino el caso límite de recibir un corazón nuevo, unos pulmones nuevos, etc. El Criterio Físico debe apelar sólo a la continuidad de mi cerebro. Entonces me referí al hecho de que muchas personas han sobrevivido con uno de los hemisferios destruido. Como está claro que estas personas sobrevivieron, el Criterio Físico debe apelar a la continuidad, pero no de todo el cerebro, sino de una cantidad de cerebro suficiente como para que sirva de base a la vida consciente.

Semejante continuidad no es lógicamente una relación uno-uno. En el caso imaginario en que me divido, cada una de las dos personas resultantes tiene lo suficiente de mi cerebro como para dar sustento a la vida consciente. Y no podemos descartar este caso con la afirmación de que nunca podría suceder. Su rasgo más inquietante, la división de la conciencia, ya ha ocurrido. Puede seguir siendo imposible dividir el cerebro inferior, pero se trata de una imposibilidad meramente técnica. De la misma manera, el teletransporte puede que nunca sea posible. Pero esa imposibilidad no debilita el argumento de Williams contra el Criterio Psicológico Amplio. Y si él apela a tales casos en este argumento, no puede descartar el caso imaginario de mi división completa.

El argumento de Williams implica que, en este caso, yo dejaré de existir, y las dos personas resultantes serán personas nuevas. Él tiene por consiguiente que revisar el Criterio Físico, de modo que adopte una forma no ramificada. Alguien podría apelar a la versión que describí. Esta es

> *El Criterio Físico*: Si va a haber una persona futura con lo suficiente de mi cerebro para que sea el cerebro de una persona viva, esta persona será yo, a no ser que vaya a haber también alguien más con bastante de mi cerebro.

Williams rechazaría este criterio porque viola sus dos requisitos.

Una vez más vale la pena dar un ejemplo. Supongamos que mi división procede como sigue. Tengo dos hermanos con el cerebro fatalmente dañado, Jack y Bill. Un cirujano primero extrae y luego divide mi cerebro. Las dos mitades se llevan entonces a diferentes alas del hospital, donde serán trasplantadas a los cuerpos de mis dos hermanos. Si apelamos al Criterio Físico, tenemos que decir lo siguiente. Supongamos que una mitad de mi cerebro se trasplanta con éxito en el cuerpo de Jack. Antes de que la otra mitad pueda ser trasplantada, se cae en el suelo de cemento. Si esto es lo que ocurre, despertaré en el cuerpo de Jack. Pero si la otra mitad fuera trasplantada con éxito, no me despertaría en ningún cuerpo.

Estas afirmaciones violan el requisito (1). Que yo sea la persona en el cuerpo de Jack debe depender sólo de los rasgos intrínsecos de la relación entre esa persona y yo. No puede pensarse con verosimilitud que dependa de lo que ocurre en el otro ala del hospital. Lo que ocurre en otra parte parece ser tan irrelevante como si los científicos de Io hacen una Réplica mía. Ocurra lo que le ocurra a Bill, y a la otra mitad de mi cerebro, mi relación con la persona del cuerpo de Jack tiene que ser la misma. El Criterio Físico niega esta afirmación. Y, comparado con la importancia del hecho de que la mitad de mi cerebro va a sobrevivir en el cuerpo de Jack, lo que le ocurre a la otra mitad es, para mí, relativamente trivial. Por eso este criterio también viola el requisito (2).

Williams podría sugerir

> *El Nuevo Criterio Físico*: Una persona futura será yo si y sólo si la persona vive y además tiene *más de la mitad* de mi cerebro [50].

Es un rasgo intrínseco de esta relación el que pueda adoptar sólo una forma uno-uno. Es lógicamente imposible para dos personas futuras que ambas tengan más de la mitad de mi cerebro. Este criterio por tanto cumple el requisito (1).

Pero es incapaz de cumplir el otro. Yo podría ser enteramente continuo desde el punto de vista psicológico con alguna persona futura *tanto* cuando esta persona tiene la mitad de mi cerebro *como*

[50] Cf. Wiggins (1), p. 55.

cuando tiene un poco más de la mitad. Y, para los que piensan que lo que importa es la continuidad física, la diferencia entre estos casos tiene que ser trivial. El segundo implica la continuidad de sólo unas pocas células más. Es un hecho trivial el que una persona futura tenga la mitad de mi cerebro, o algo más que la mitad. El Nuevo Criterio Físico, por tanto, viola el requisito (2).

Hay otra objeción contra este criterio. Alguien podría sufrir lesiones que causaran que más de la mitad de su cerebro dejara de funcionar. Aunque esa persona tendría paralizada más de la mitad de su cuerpo, y podría necesitar que la colocaran en un corazón y un pulmón artificiales, su vida mental podría seguir inalterada. Menos de la mitad de un cerebro podría ser suficiente para proporcionar continuidad psicológica completa. Naturalmente, pensaríamos que esa persona sobrevive a sus lesiones. Pero, según el Nuevo Criterio Físico, tenemos que afirmar que la persona deja de existir. La persona que hay en ese cuerpo es alguien diferente, una persona nueva que es simplemente exacta a la anterior. Esto resulta difícil de creer. Es una segunda objeción poderosa contra este criterio.

En todas sus versiones posibles, el Criterio Físico afronta fuertes objeciones. Y las hay similares contra el Criterio Psicológico. Los requisitos de Williams son ambos plausibles. Hemos encontrado que *no hay ningún criterio de identidad plausible que pueda cumplirlos los dos*. (Si fuéramos entidades que existen separadamente, como Egos Cartesianos, nuestro criterio podría cumplir estos requisitos; pero tenemos razones suficientes para rechazar esta concepción.)

Volvamos ahora a la Concepción Reduccionista. Reconsideremos el caso en que la mitad de mi cerebro se trasplanta con éxito al cuerpo de Jack. ¿Cuál es mi relación con la persona que se despierta en el cuerpo de Jack? Esta relación es la de continuidad psicológica, con su causa normal, la existencia continua de suficiente cantidad de cerebro. Hay también mucho parecido físico. Como reduccionista, afirmo que mi relación con la persona en el cuerpo de Jack contiene lo que fundamentalmente importa. Esta afirmación sigue en pie ocurra lo que ocurra con otras personas en otra parte. Con una revisión, mi concepción cumple el primer requisito de

Williams. Él dice que si yo voy a ser una persona futura debe depender sólo de mi relación con esa persona futura. Yo hago una afirmación similar. En vez de preguntar si voy a ser una persona futura, pregunto si mi relación con esta persona contiene lo que importa. Como Williams, yo puedo afirmar que la respuesta tiene que depender sólo de los rasgos *intrínsecos* de mi relación con esta persona futura.

La Concepción Reduccionista puede cumplir esta versión revisada del requisito (1). Supongamos que la otra operación tiene éxito. Alguien se despierta en el cuerpo de Bill. Según mi modo de ver, esto no cambia la relación que mantengo con la persona en el cuerpo de Jack. Y supone como mucho una pequeña diferencia para su importancia. Esta relación todavía contiene lo que fundamentalmente importa. Como se da ahora de una forma ramificada, nos vemos forzados a cambiar su *nombre*. No podemos llamar a cada una de sus ramas identidad personal. Pero este cambio en el nombre de la relación no tiene importancia.

Esta Concepción Reduccionista también cumple el análogo del requisito (2). Los juicios de identidad personal tienen gran importancia. Por eso afirma Williams que no deberíamos hacer uno de estos juicios y negar otro sin que haya una diferencia importante en nuestras razones. Según esta Concepción Reduccionista, deberíamos coger la importancia que damos a los juicios de identidad y dársela a una relación diferente. Según esta concepción, lo que es importante es la relación R: conexividad psicológica y/o continuidad, con la clase correcta de causa. A diferencia de la identidad, esta relación no puede dejar de darse por causa de una diferencia trivial en los hechos. Si esta relación deja de darse, hay una profunda diferencia en los hechos. Esto cumple el requisito (2).

En el caso en que me divido, aunque mi relación con cada una de las personas resultantes no puede llamarse identidad, contiene lo que fundamentalmente importa. Cuando aquí negamos la identidad, no tenemos necesidad de negar un juicio importante. Como mi relación con cada una de las personas resultantes es casi tan buena como si fuera identidad, puede conllevar la mayoría de las implicaciones corrientes de la identidad. Así, podría decirse que, aunque la persona en el cuerpo de Jack no pueda llamarse yo, porque el otro trasplante tiene éxito, puede merecer en la misma medida que yo castigo o recompensa por lo que yo haya hecho. Igual que la persona en el cuerpo de Bill. Como Wiggins escribe: «un malhechor difícilmente podría eludir su responsabilidad planificando su propia fisión» [51].

Hay aquí preguntas que responder. Si al malhechor se le sentencia a veinte años de cárcel, ¿cada persona resultante debería cumplir veinte años o sólo diez? Discuto algunas de estas cuestiones en el capítulo 15. Pero no ponen en duda la afirmación general que he hecho. Si aceptamos la Concepción Reduccionista, es R y no la identidad lo que importa.

Puede pensarse que, si esto es así, debemos dar a R la importancia que ahora damos a la identidad personal. Esto no se sigue. Si creemos que la identidad personal tiene una gran importancia, esto puede ser porque creemos en la Concepción No-Reduccionista. Si cambiamos de opinión, y nos hacemos reduccionistas, podemos también cambiar de opinión sobre la importancia de la identidad personal. Podemos aceptar que la relación R tiene casi todo lo que de importancia tiene, *según la Concepción Reduccionista*, la identidad personal. Y podemos aceptar que, según esta concepción, lo que fundamentalmente importa no es la identidad personal sino R. Pero podemos pensar que estas dos relaciones tienen mucha *menos* importancia que la que *tendría* la identidad personal si fuera verdadera la Concepción No-Reduccionista. Discuto esta creencia en los capítulos 14 y 15.

Esta creencia no afecta a mis afirmaciones acerca de los requisitos de Williams. Si asumimos que la identidad es lo que importa, no podemos cumplir estos requisitos. Como debemos rechazar la Concepción No-Reduccionista, nuestro criterio de identidad debería ser o el Físico o el Psicológico. Y, como he sostenido, no hay

[51] Wiggins (4), p. 146. Como veremos abajo, algunos autores rechazan esta afirmación.

ninguna versión plausible de estos criterios que cumpla los dos requisitos de Williams.

Añado estas observaciones. Ahora que hemos visto que la identidad no es lo que importa, no deberíamos tratar de revisar ni de ampliar nuestro criterio de identidad para que coincida más a menudo con lo que importa. Según cualquier entendimiento natural de la identidad personal, tal coincidencia podría ser sólo parcial, como muestra el caso de la división. Y revisar nuestro criterio puede sugerir, engañosamente, que la identidad es lo que importa.

Williams da otra razón para exigir que el criterio de identidad sea lógicamente de la forma uno-uno, y de un modo que no sea arbitrario. «A no ser que haya un requisito semejante, no puedo ver cómo uno va a preservar y a explicar la verdad evidente de que los conceptos de identidad y de similitud exacta son conceptos diferentes» [52].

La concepción reduccionista preserva y explica esta verdad. He descrito casos en que hay dos personas que son exactamente iguales pero no son numéricamente idénticas. Esto puede ser lo que ocurra en el caso de la línea secundaria, en donde yo hablo con mi Réplica. Por eso comprendemos la pregunta, «¿Es la misma persona que yo o meramente otra persona exactamente igual?». He afirmado que en algunos casos, como los que están en el medio del Espectro Físico, no hay una diferencia real entre que la persona resultante sea yo y que sea alguien diferente que es exactamente igual que yo. La Concepción Reduccionista implica que, en algunos casos, no hay una diferencia real entre la identidad numérica y la semejanza exacta. Pero como reconoce otros casos donde esto es una diferencia real, preserva y explica la verdad de que son diferentes conceptos.

Discutí las manifestaciones de Williams para ver si aportan una tercera concepción, diferente tanto de la Concepción Reduccionista como de la de que somos entidades que existen separadamente. Concluyo que estas manifestaciones no aportan una tercera concepción tal. Lo que aportan son razones adicionales para aceptar la concepción reduccionista. Los requisitos de Williams son ambos plausibles. Si creemos que la identidad es lo que importa, no podemos cumplir estos requisitos. Pero si aceptamos la Concepción Reduccionista y apelamos a la relación R, *podemos* cumplir los requisitos análogos.

[52] Williams (9), reeditado en Williams (2), p. 24.

92. Wittgenstein y Buda

Wittgenstein habría rechazado la Concepción Reduccionista. Él pensaba que nuestros conceptos dependen del darse de determinados hechos, y que no deberíamos considerar casos imaginarios en que estos hechos ya no se dan. Los argumentos a favor de la Concepción Reduccionista apelan a tales casos.

Este desacuerdo es sólo parcial. La mayor parte de la gente tiene creencias acerca de estos casos imaginarios. Como he defendido, estas creencias implican que somos entidades que existen separadamente, distintas de nuestros cerebros y de nuestros cuerpos, y entidades cuya existencia tiene que ser todo-o-nada. Una afirmación reduccionista central es que deberíamos rechazar estas creencias. *Wittgenstein habría estado de acuerdo*. Dado su acuerdo sobre esta afirmación, no necesito discutir la concepción de Wittgenstein, ni algunas otras concepciones parecidas, como la que fue avanzada por Wiggins [53].

Con dos excepciones que pronto mencionaré, creo que he considerado ahora las concepciones que necesitan ser consideradas en este debate. Y no he considerado visiones mantenidas en épocas o civilizaciones diferentes. Este hecho sugiere una inquietante posibilidad. Yo creo que mis afirmaciones se aplican a todas las personas, en todo tiempo. Sería preocupante descubrir que son sólo parte de una línea de pensamiento en la cultura de la Europa y América modernas.

[53] En Wiggins (3). Por desgracia, Wiggins no continúa aquí la discusión de su caso imaginario de división en Wiggins (1).

Afortunadamente, esto no es verdad. Afirmo que, cuando preguntamos qué son las personas y cómo continúan existiendo, la cuestión fundamental es una elección entre dos concepciones. Según una de ellas, somos entidades que existen separadamente, distintas de nuestros cerebros, cuerpos y experiencias, y entidades cuya existencia tiene que ser todo-o-nada. La otra concepción es la reduccionista. Y sostengo que, de las dos, es la segunda la que es verdadera. Como muestra el Apéndice J, *Buda habría estado de acuerdo.* La Concepción Reduccionista no es sólo parte de una tradición cultural. Como he dicho, puede ser la concepción verdadera acerca de todas las personas en todas las épocas.

93. ¿SOY ESENCIALMENTE MI CEREBRO?

Nagel sugiere una idea de un tipo que no he discutido. Sugiere que lo que realmente soy es lo que quiera que sea la causa de mi continuidad psicológica. Dado lo que ahora sabemos, lo que realmente soy es mi cerebro. Según esta concepción, además, soy mi cerebro *esencialmente.* No puedo decidir adoptar una concepción diferente acerca de mí mismo.

Nagel apoya esta idea de tres formas. Da dos argumentos, que trato de contestar en el Apéndice D. También afirma que su idea es intuitivamente plausible. Como él mismo escribe, el cerebro

> «me parece que es algo sin lo cual yo no podría sobrevivir —de manera que, si fuera producida una réplica mía físicamente distinta que fuese psicológicamente continua conmigo aunque mi cerebro hubiera sido destruido, esta réplica no sería yo y su supervivencia no sería tan buena como mi supervivencia».

Nagel está considerando aquí un caso como el del teletransporte. El escáner replicador podría tener su utilidad aquí en la Tierra. Lo podría usar yo como una manera más segura de atravesar Manhattan, o como cura en caso de que, mientras cruzo Manhattan andando, fuese mortalmente apuñalado. Como precaución, antes de cada paseo, podría dejar guardada un nuevo cianotipo mío.

Nagel sugiere que el teletransporte no es sólo que no sea tan bueno como la supervivencia corriente, sino que es casi tan malo como la muerte. Después de describir su caso imaginario, escribe,

> «no sobreviviré a la noche... la réplica no será yo. Tratando de armarme de valor, me preparo para el final».

Esto sugiere que, según el modo de ver de Nagel, es la identidad personal lo que importa. Admito que, en un caso como el del teletransporte, muchos aceptarían este modo de ver las cosas. Pensarían que lo que importa es la supervivencia del cerebro. En el Apéndice D describo dos casos en que esto es más difícil de creer.

Como la idea de Nagel es, en algunos casos, intuitivamente plausible, y el Apéndice D tal vez fracase a la hora de responder a sus argumentos, voy a dar en la Sección 98 una clase diferente de respuesta a esta idea y otras similares.

Hay otra concepción que debería considerar. Explico en el Apéndice E cómo, no obstante un aparente desacuerdo, la opinión de Nozick es una versión del Reduccionismo.

94. ¿ES CREÍBLE LA CONCEPCIÓN VERDADERA?

Nagel afirmó una vez que, aunque sea verdadera la Concepción Reduccionista, a nosotros nos resulta psicológicamente imposible creer en ella. Por ello revisaré brevemente los argumentos que di arriba. Después preguntaré si *yo* puedo decir honestamente que creo en mis conclusiones. Si puedo, daré por sentado que no soy único. Habría al menos algunas otras personas que pueden creer la verdad.

Primero matizaré mi afirmación de que he descrito la concepción verdadera. Es difícil explicar con exactitud lo que sostiene un reduccionista. Y es difícil explicar con exactitud lo que está implicado en la identidad a través del tiempo. Por eso es probable que al describir la Concepción Reduccionista acerca de la identidad haya cometido errores.

Puede que esos errores no arruinen completamente mis argumentos. Wittgenstein sugirió esta analogía. Supongamos que estoy cambiando de sitio los libros de una biblioteca. Al comienzo de este cambio, coloco dos libros uno al lado del otro en un estante determinado, porque esos libros deben ir juntos. Puede que yo sepa que, al final del cambio de lugar, esos dos libros van a estar en un estante diferente. Pero aún así todavía vale la pena colocarlos juntos ahora. Si deben estar juntos, todavía estarán juntos después del cambio de sitio.

Afirmo que una persona no es como un Ego Cartesiano, un ser cuya existencia tenga que ser todo-o-nada. Una persona es como una nación. En la verdadera explicación de la identidad a través del tiempo, estas dos clases de entidades van juntas. Son como los dos libros que empiezo colocando en un estante. En mis afirmaciones sobre el Reduccionismo y la identidad, puedo haber cometido errores. Esto sería como el hecho de que los dos libros deberían ir en un estante diferente. Pero mi afirmación principal es que las personas son como las naciones, no como Egos Cartesianos. Si esta afirmación es verdadera, no se vería arruinada por mis errores. En la explicación que no comete errores, las personas y las naciones todavía irían juntas. La explicación de su identidad a través del tiempo sería similar, en sus rasgos esenciales.

Distinguí dos opiniones sobre la naturaleza de las personas. Según la Concepción No-Reduccionista, una persona es una entidad que existe separadamente, distinta de su cerebro y de su cuerpo y sus experiencias. Según la versión mejor conocida de esta concepción, una persona es un Ego Cartesiano. Según la Concepción Reduccionista que yo defiendo, las personas existen. Y una persona es distinta de su cerebro y de su cuerpo y sus experiencias. Pero las personas no son entidades que existan separadamente. La existencia de una persona, durante un período dado, consiste justo en la existencia de su cerebro y de su cuerpo, y en el pensar de sus pensamientos y en el hacer de sus acciones, y en la ocurrencia de muchos otros sucesos físicos y mentales.

Como estas concepciones están en desacuerdo en lo que respecta a la naturaleza de las personas, también lo están en lo que hace a la naturaleza de la identidad personal a través del tiempo. Según la Concepción Reduccionista, la identidad personal nada más que implica la continuidad física y psicológica. Como he defendido, las dos pueden describirse de un modo impersonal. Estas dos clases de continuidad pueden describirse sin afirmar que las experiencias son tenidas por una persona. Un reduccionista también afirma que la identidad personal no es lo que importa. La identidad personal nada más que implica ciertas clases de conexividad y continuidad, cuando estas se dan de una forma uno-uno. Estas relaciones son lo que importa.

Según la Concepción No-Reduccionista, la identidad personal es lo que importa. Y no implica sólo continuidad física y psicológica. Es un hecho adicional separado, que, en cualquier caso, tiene que darse del todo o no darse en absoluto. La unidad psicológica se explica por la propiedad. La unidad de la conciencia en un momento dado se explica por el hecho de que varias experiencias están siendo tenidas por una persona. Y la unidad de la vida de una persona se explica de la misma manera. Estas diversas afirmaciones, como he sostenido, tienen que mantenerse o caer juntas.

Concedí que la Concepción No-Reduccionista podría haber sido verdadera. Podría, por ejemplo, haber habido evidencia que apoyara la creencia en la reencarnación. Pero de hecho no hay buena evidencia a favor de esta concepción, y sí mucha en su contra.

Parte de la evidencia la proporcionan los casos reales de mentes divididas. Como han sido desconectados sus hemisferios, hay personas que tienen dos corrientes de conciencia, en cada una de las cuales son inconscientes de la otra. Podríamos decir que, en un caso así, hay dos personas diferentes en el mismo cuerpo. Esto sería tratar a esos casos como si fueran parecidos al caso imaginario en que me divido, que reviso abajo. Nuestra alternativa es afirmar, en relación con estos casos reales, que hay una única persona con dos corrientes de conciencia.

Si hacemos esta afirmación, ¿cómo podemos explicar la unidad de la conciencia en cada corriente? No la podemos explicar diciendo que las diversas experiencias diferentes de cada corriente están siendo tenidas por la misma persona o sujeto de experiencias. Esto

describe las dos corrientes como si fueran una. Si pensamos que la unidad de la conciencia tiene que explicarse adscribiendo diferentes experiencias a un sujeto particular, tenemos que afirmar que en estos casos, aunque haya sólo una persona, hay dos sujetos de experiencias. De esta forma, tenemos que decir que puede haber en la vida de una persona sujetos de experiencias que *no* son personas. Es difícil de creer que haya realmente algo semejante. Estos casos se explican mejor por el criterio psicológico reduccionista, que afirma que, en un momento dado, hay un estado de conciencia de las experiencias en una corriente de conciencia, y otro estado de conciencia de las experiencias en la otra corriente.

Aunque le plantean este problema a la Concepción No-Reduccionista, estos casos de mentes divididas son sólo una pequeña parte de la evidencia contra esta concepción. No hay evidencia de que el portador de la continuidad psicológica sea algo cuya existencia, como la de un Ego Cartesiano, tenga que ser todo-o-nada. Y hay mucha evidencia a favor de que nuestros rasgos psicológicos dependen de estados y sucesos que tienen lugar en el cerebro. La existencia continua de un cerebro no tiene por qué ser del tipo todo-o-nada. La conexividad física puede ser una cuestión de grado. Y hay un sinnúmero de casos reales en los que la conexividad psicológica se da sólo de ciertas formas, o en un grado reducido.

Tenemos evidencia suficiente para rechazar la Concepción No-Reduccionista. La Concepción Reduccionista es, como digo, la única alternativa. Consideré posibles terceras posiciones, y no encontré ninguna que fuera no-reduccionista y al mismo tiempo una que tuviéramos suficientes razones para aceptar. Más exactamente, aunque estas otras concepciones difieran de otras maneras, las plausibles no niegan una afirmación central del reduccionista: están de acuerdo en que no somos entidades que existen separadamente, distintas de nuestros cerebros y cuerpos y nuestras experiencias, cuya existencia tiene que ser del tipo todo-o-nada.

Además de hacer estas declaraciones acerca de los hechos, las hice sobre nuestras creencias. Nos enteramos de que tenemos estas creencias cuando consideramos ciertos casos imaginarios.

Uno de ellos es el caso en que, manipulando mi cerebro, un cirujano elimina gradualmente toda mi continuidad psicológica. Describí otras tres gamas de casos. En los Espectros Psicológico y Físico habría, entre yo y una persona futura, todos los grados posibles de conexividad psicológica o de conexividad física. En el Espectro Combinado, habría todos los grados posibles de los dos tipos de conexividad.

Como la mayoría de nosotros, tengo la poderosa inclinación a pensar que cualquier persona futura tiene que ser o yo o alguien distinto. Y estoy inclinado a creer que hay siempre una profunda diferencia entre estos dos resultados. Como no soy una entidad que exista separadamente, estas creencias no pueden ser verdaderas. Esto se ve perfectamente considerando el Espectro Combinado. En el caso del extremo cercano, donde no me harían nada, la persona resultante sería ciertamente yo. En el caso del extremo lejano, donde no habría conexiones entre la persona resultante y yo, esta persona sería ciertamente alguien distinto. Si cualquier persona futura tiene que ser o yo o alguien distinto, tiene que haber una línea en esta gama de casos hasta la cual la persona resultante sería yo, y más allá de la cual sería alguien distinto. Si siempre hay una profunda diferencia entre que una persona sea yo y que sea alguien distinto, tiene que haber una profunda diferencia entre estos dos casos de este espectro. Tiene que haber una profunda diferencia tal, aunque nunca pudiésemos descubrir dónde llega. Estas afirmaciones son falsas. No hay una profunda diferencia entre ningún caso vecino en esta gama. Las únicas diferencias son que, en uno de los casos, el cirujano reemplazaría unas cuantas células más, y provocaría un cambio psicológico más pequeño.

Cuando considero esta gama de casos, me veo forzado a abandonar al menos una de las dos creencias mencionadas arriba. No puedo seguir creyendo que cualquier persona futura tiene que ser yo o alguien distinto, y que siempre hay, además, una profunda diferencia entre estos dos resultados.

También me veo forzado a aceptar otra parte de la Concepción Reduccionista. Supongamos que estoy a punto de sufrir una de las operaciones en la mitad de este espectro. Sé que, entre la persona

resultante y yo, habrá ciertas clases y ciertas cantidades de conexividad física y psicológica. Según la Concepción Reduccionista, conociendo estos hechos conozco toda la verdad de lo que va a ocurrir. Cuando esté a punto de perder la conciencia, puedo preguntar, «¿Estoy a punto de morir? ¿O seré la persona resultante?». Y estoy inclinado a creer que estas son siempre dos posibilidades diferentes, de las que una tiene que ser la verdad. Según la Concepción Reduccionista, aquí lo que hay es una cuestión vacía. A veces hay una diferencia real entre que una persona futura sea yo y que sea alguien distinto. Pero no hay esa diferencia real en los casos de la mitad del Espectro Combinado. ¿Cuál podría ser la diferencia? ¿Qué es lo que podría hacer verdadero que la persona resultante fuese yo o que fuese alguien distinto? Como no soy una entidad que exista separadamente, no hay nada que pudiera realizar estas diferentes posibilidades, de las que cada una podría ser verdadera. En estos casos, podríamos decir que la persona resultante será yo, o bien podríamos decir que yo moriré y ella será alguien distinto. Pero estos no son aquí resultados diferentes. Son sólo diferentes descripciones del mismo resultado.

Para ilustrar estas afirmaciones repetí la comparación de Hume. Las personas con como naciones, clubes o partidos políticos. Si consideramos estas otras entidades, la mayoría de nosotros aceptará una Concepción Reduccionista. Recordemos el partido político que se escindió, y se convirtió en dos partidos rivales. Podemos preguntar, «¿Dejó de existir el partido original, o siguió existiendo como el uno o el otro de los partidos resultantes?». Pero no creemos que esto sea una pregunta real sobre posibilidades diferentes, una de las cuales tenga que ser lo que ocurrió. Esta pregunta es una pregunta vacía. Aunque no tengamos respuesta para ella, podríamos conocer toda la verdad de lo que ocurrió.

Como aceptamos la Concepción Reduccionista por lo que respeta a partidos políticos, clubes y naciones, comprendemos más o menos lo que afirma la Concepción Reduccionista cuando se aplica a personas. Pero la mayor parte de nosotros está fuertemente inclinada a rechazar esta concepción. Estamos fuertemente inclinados a creer que tiene que haber siempre una diferencia entre que una persona futura sea yo y que sea alguien distinto. Considerar mi Espectro Combinado puede que no sea suficiente para persuadirnos de hacernos reduccionistas. Por eso di más argumentos.

Un argumento se refería al caso imaginario en que me divido. Podría afirmarse que es absolutamente imposible la división de una corriente de conciencia. Pero lo que sucede tiene que ser posible, y en las vidas de varias personas esto ha sucedido. Mi división imaginaria es una ampliación de estos casos reales.

En este caso imaginario, cada mitad de mi cerebro se trasplanta con éxito a otro cuerpo. ¿Qué es lo que me ocurre? A no ser que distorsionemos de maneta grotesca el concepto de persona, las únicas respuestas posibles son que yo seré una de las personas resultantes, o bien la otra, o ninguna de las dos. Si pensamos que la identidad es lo que importa, cada respuesta de estas es difícil de aceptar. Dada la exacta semejanza de las dos personas resultantes, es difícil de creer que yo seré *una* de estas dos personas. Si no voy a ser *ninguna* de las personas, y la identidad es lo que importa, debo considerar la división como equivalente a la muerte. Pero esto también es difícil de creer. Mi relación con cada una de las personas resultantes contiene todo lo que se necesitaría para la supervivencia. Esta relación no puede llamarse identidad porque, y sólo porque, se da entre yo y *dos* personas futuras. En la muerte corriente, la relación se da entre yo y *ninguna* persona futura. Aunque la doble supervivencia no puede describirse en el lenguaje de la identidad, no es equivalente a la muerte. Dos no es igual a cero.

Este caso imaginario apoya otra parte de la Concepción Reduccionista. No sólo es cada una de las respuestas posibles difícil de creer. Es difícil ver cómo el caso implicaría diferentes posibilidades, cualquiera de las cuales podría ser la verdad. Si no soy un Ego Cartesiano, ¿qué podría hacer verdadero que yo fuese una u otra de las dos personas resultantes? Si estamos ante posibilidades diferentes, ¿en qué podría consistir la diferencia? Parece que no hay respuesta a esta pregunta. Cada una de las personas resultantes tendrá la mitad de mi cerebro, y será completamente continua conmigo desde el punto de vista psicológico. Parecemos forzados a concluir

que esta es una descripción completa del caso. Comprendemos la pregunta, «¿Seré una u otra o ninguna de las dos personas?». Pero esta es otra pregunta vacía. No hay aquí diferentes posibilidades, de las que una tenga que ser verdadera. Simplemente se trata de diferentes descripciones del mismo resultado.

La mejor descripción es que no seré ninguna persona resultante. Pero esto no implica que deba considerar la división casi tan mala como la muerte. Como defendí, debo considerarla casi tan buena como la supervivencia corriente. Para algunas personas, sería algo mejor; para otras, sería algo peor. Como no puedo ser la misma persona que las dos personas resultantes, pero mi relación con cada una de ellas contiene lo que fundamentalmente importa en la supervivencia ordinaria, el caso muestra que la identidad no es lo que importa. Lo que importa es la relación R: conexividad psicológica y/o continuidad psicológica con la clase correcta de causa.

Ahora he pasado revista a los principales argumentos a favor de la Concepción Reduccionista. ¿Encuentro imposible creer en esta concepción?

Lo que encuentro es esto. Puedo creer en esta concepción al nivel intelectual o reflexivo. Me convencen los argumentos a su favor. Pero pienso que es probable que, a otro nivel, siempre vaya a tener dudas.

Mi creencia es más firme cuando considero algunos casos imaginarios. Estoy convencido de que, si me dividiera, sería una pregunta vacía la de si seré una u otra o ninguna de las personas resultantes. Creo que no hay nada que pudiera realizar estas posibilidades diferentes, si pensamos que cualquiera de las cuales podría ser lo que realmente ocurriría. Y estoy convencido de que, en los casos centrales del tercer espectro, es una pregunta vacía la de si la persona resultante sería yo.

Cuando considero otros casos concretos, mi convicción es menos firme. Un ejemplo es el teletransporte. Imagino que me encuentro en el cubículo, a punto de apretar el botón verde. Podría tener dudas, de repente. Podría tener la tentación de cambiar de opinión y pagar el billete, mucho más caro, de un viaje en nave espacial.

Sospecho que revisar mis argumentos nunca eliminaría completamente mis dudas. Al nivel intelectual o reflexivo, seguiría convencido de que la Concepción Reduccionista es la verdadera. Pero a un nivel más elemental todavía estaría inclinado a pensar que siempre tiene que haber una diferencia real entre que una persona futura sea yo y que sea alguien distinto. Algo parecido ocurre cuando miro por una ventana en el punto más elevado de un rascacielos. Sé que no corro ningún peligro. Pero mirar hacia abajo desde esta altura de vértigo me da miedo. Tendría un miedo irracional parecido si estuviera a punto de apretar el botón verde.

Puede ser útil añadir algunos comentarios. Según la Concepción Reduccionista, mi existencia continua nada más que implica continuidad física y psicológica. Según la Concepción No-Reduccionista, implica un hecho adicional. Es natural creer en este hecho adicional, y creer que, comparado con las continuidades, es un hecho *profundo*, y el que realmente importa. Cuando temo que, en el teletransporte, *yo* no llegaré a Marte, lo que temo es que la causa anormal no será capaz de producir este hecho adicional. Como he defendido, no hay tal hecho. Lo que temo no ocurrirá, *nunca* ocurre. Quiero que la persona en Marte sea yo de un modo especialmente íntimo, un modo en el cual ninguna persona futura será jamás yo. Mi existencia continua nunca implica este hecho adicional profundo. Lo que temo que vaya a faltar falta *siempre*. Ni siquiera un viaje en nave espacial produciría el hecho adicional en el que estoy inclinado a creer.

Cuando llego a ver que mi existencia continua no implica este hecho adicional, pierdo la razón que tenía para preferir un viaje en nave espacial. Pero si lo juzgamos desde el punto de vista de mi creencia anterior, esto no ocurre porque el teletransporte sea *casi tan bueno como* la supervivencia ordinaria. Ocurre porque la supervivencia ordinaria es *casi tan mala como*, o un poco mejor que, el teletransporte. *La supervivencia ordinaria es casi tan mala como ser destruido y replicado.*

Repitiendo argumentos como estos podría conseguir reducir el miedo que siento. Llegaría a ser capaz de determinarme a apretar el

botón verde. Pero me parece que nunca perderé completamente mi creencia intuitiva en la Concepción No-Reduccionista. Es difícil confiar serenamente en mis conclusiones reduccionistas. Es difícil creer que la identidad personal no es lo que importa. Si mañana alguien va a sufrir dolores horrorosos, es difícil de creer que podría ser una pregunta vacía la de si *yo* voy a ser el que los sienta. Y es difícil de creer que, si yo estoy a punto de perder la conciencia, podría no haber respuesta a la pregunta «¿Estoy a punto de morir?».

Nagel dijo una vez que es psicológicamente imposible creer en la Concepción Reduccionista. Buda dijo que, aunque es muy difícil, es posible. Tras pasar revista a mis argumentos, encuentro que, en el nivel intelectual o reflexivo, aunque es muy difícil creer en la concepción reduccionista, es posible hacerlo. Las dudas o los miedos que me quedan me parecen irracionales. Como puedo creer en esta concepción, doy por sentado que hay otros que también pueden hacerlo. Podemos creer la verdad sobre nosotros mismos.

13

LO QUE IMPORTA

95. Liberación del yo

La verdad es muy diferente de lo que estamos inclinados a creer. Aunque no lo sepamos, la mayoría de nosotros es no-reduccionista. Si consideramos mis casos imaginarios, estaríamos fuertemente inclinados a creer que nuestra existencia continua es un hecho adicional profundo, distinto de la continuidad física y psicológica, y un hecho que tiene que ser todo-o-nada. Nada de esto es verdad.

¿Es deprimente la verdad? Algunos la encontrarán deprimente. Pero yo la encuentro liberadora y consoladora. Cuando yo creía que mi existencia era tal hecho adicional, me daba la impresión de estar prisionero dentro de mí mismo. Mi vida parecía un túnel de cristal por el que me movía más rápido cada año y al final del cual no había sino oscuridad. Cuando cambié de manera de pensar, las paredes de mi túnel de cristal desaparecieron. Ahora vivo al aire libre. Hay todavía una diferencia entre mi vida y las vidas de otras personas. Pero la diferencia es menor. Las otras personas están más próximas. Y yo estoy menos preocupado por el resto de mi propia vida, y más preocupado por la vida de los demás.

Cuando creía en la Concepción No-Reduccionista, también me preocupaba más mi muerte inevitable. Después de mi muerte no habrá nadie viviendo que vaya a ser yo. Ahora puedo redescribir este hecho. Aunque después habrá muchas experiencias, ninguna de ellas estará conectada a mis experiencias presentes por cadenas de conexiones directas como las que están implicadas en la memoria experiencial, o en la realización de una intención anterior. Algunas de estas experiencias futuras pueden estar relacionadas con mis experiencias presentes de modos menos directos. Habrá después algunos recuerdos de mi vida. Y puede haber después pensamientos influidos por los míos, o cosas hechas como resultado de mis consejos. Mi muerte romperá las relaciones más directas entre mis experiencias presentes y las experiencias futuras, pero no romperá otras variadas relaciones. Esto es todo lo que hay en el hecho de que no habrá nadie viviendo que sea yo. Ahora que he visto esto, mi muerte me parece menos mala.

En vez de decir, «Estaré muerto», debería decir, «No habrá experiencias futuras que vayan a estar relacionadas, de ciertos modos, con estas experiencias presentes». Como me recuerda lo que este hecho conlleva, esta redescripción lo hace menos deprimente. Supongamos a renglón seguido que tengo que sufrir una terrible experiencia. En vez de decir, «La persona que va a sufrir será yo», yo debería decir, «Habrá un sufrimiento que estará relacionado, de ciertos modos, con estas experiencias presentes». Una vez más, el hecho redescrito me parece menos malo.

Puedo incrementar estos efectos imaginando vívidamente que estoy a punto de sufrir una de las operaciones que he descrito. Me imagino que estoy en un caso central del Espectro Combinado, donde es una pregunta vacía la de si estoy a punto de morir. Es muy difícil creer que esta podría ser una pregunta vacía. Cuando repaso los argumentos a favor de esta creencia, y me vuelvo a convencer, esto por un momento deja en suspenso mi natural preocupación por mi futuro. Cuando mi futuro real vaya a ser muy negro —como ocurriría si fuese a ser torturado, o tuviera que enfrentarme al amanecer al pelotón de ejecución— será bueno tener este modo de dejar en suspenso brevemente mi preocupación.

Tras meditar una y otra vez estos argumentos, Hume vino a parar a «la más deplorable condición que se pueda imaginar, rodeado por la más profunda oscuridad» [54]. La cura consistió en cenar y jugar al backgammon con sus amigos. Los argumentos de Hume dieron pie al escepticismo absoluto. Por eso trajeron oscuridad y soledad total. Los argumentos a favor del Reduccionismo tienen en mí el efecto contrario. Dar vueltas a estos argumentos elimina el muro de cristal entre los demás y yo. Y, como he dicho, me importa menos mi muerte. Esta no es nada más que el hecho de que, tras un tiempo determinado, ninguna de las experiencias que ocurrirán estará relacionada, de ciertos modos, con mis experiencias presentes. ¿Acaso esto importa tanto?

96. La continuidad del cuerpo

Estoy contento de que la Concepción Reduccionista sea verdadera, ya que afecta a mis emociones de esta manera. Pero se trata sólo de un informe de sus efectos psicológicos. Los efectos en otras personas bien pueden ser diferentes.

Hay otras diversas cuestiones que quedan por discutir. Y al discutirlas puedo hacer algo más que informar de hechos acerca de mis reacciones. Las respuestas a estas preguntas dependen en parte de la fuerza de ciertos argumentos. Primero discutiré lo que, como reduccionistas, debemos afirmar que es lo que importa. Después preguntaré cómo, si hemos cambiado de modo de pensar en lo que respecta a la naturaleza de la identidad personal, debemos cambiar nuestras creencias sobre la racionalidad y la moralidad.

Como demuestra el caso de mi división, la identidad personal no es lo que importa. Simplemente ocurre que, en la mayoría de los casos, la identidad personal coincide con lo que importa. ¿Qué es lo que importa *de la manera en que* se piensa equivocadamente que la identidad personal importa? ¿De qué es racional preocuparse en nuestra preocupación por nuestro propio futuro?

[54] Hume (1), p. 269.

Esta pregunta puede reformularse. Asumamos, por simplicidad, que pudiera ser racional estar preocupado sólo por el propio interés de uno mismo. Supongamos que soy egoísta, y que podría estar relacionado de diversas maneras con alguna persona resultante. ¿Cuál es la relación que justificaría la preocupación egoísta por esta persona resultante? Si el resto de la vida de esta persona va a valer la pena vivirse, sin ninguna duda, ¿de qué modo querría yo estar relacionado con ella? Si el resto de su vida va a ser mucho peor que nada, ¿de qué modo querría yo *no* estar relacionado con ella? En pocas palabras, ¿cuál es la relación que, para un egoísta, importaría fundamentalmente? Esta relación también será la que, para todos nosotros, importaría fundamentalmente en nuestra preocupación por nuestro propio futuro. Pero como podemos estar preocupados por el destino de la persona resultante, *sea cual sea* su relación con nosotros, lo más claro es preguntar qué importaría para un egoísta.

Aquí tenemos las respuestas más simples:

(1) La continuidad física,
(2) La relación R con su causa normal,
(3) R con cualquier causa fiable,
(4) R con cualquier causa.

R es conexividad y/o continuidad psicológica, con la clase correcta de causa. Si decidimos que R es lo que importa, entonces tenemos que considerar la importancia relativa de la conexividad y la continuidad. Podría sugerirse que lo que importa es *tanto* R *como* la continuidad física. Pero esto es lo mismo que la respuesta (2), puesto que la continuidad física es parte de la causa normal de R.

¿Podemos defender (1), la afirmación de que sólo importa la continuidad física? ¿Podemos decir que, si voy a ser físicamente continuo con una persona resultante, esto es lo que importa, aunque no vaya a estar R-relacionado con ella?

Volvamos a considerar el ejemplo de Williams, aquel en el que el cirujano destruye totalmente cualquier tipo distintivo de continuidad psicológica. Supongamos que este cirujano está a punto de operarme de una manera indolora, y que la persona resultante tendrá una vida que es mucho peor que nada. Si soy egoísta, podría

considerar que esta perspectiva no es peor que una muerte indolora, desde el momento en que a mí no me importa lo que le vaya a ocurrir a la persona resultante. Pero podría considerar, en cambio, que esta perspectiva es mucho peor que la muerte, porque estoy egoístamente preocupado por el atroz futuro de esta persona. ¿Cuál debería ser mi actitud?

Estaría egoístamente preocupado por el futuro de esta persona si pudiera creer justificadamente que ella va a ser *yo*, en vez de *alguien distinto* que simplemente es continuo conmigo desde el punto de vista físico. Pero, como he sostenido, esta creencia no está justificada. El ejemplo de Williams se encuentra en el extremo lejano del Espectro Psicológico. Tanto en los casos centrales de este espectro como en el extremo lejano, no hay una diferencia real entre que la persona resultante sea yo y que sea alguien distinto. No son dos posibilidades diferentes, de las que una tenga que ser verdadera. En el ejemplo de Williams, estos son todos los hechos. La persona resultante será físicamente, pero no psicológicamente, continua conmigo. Podríamos llamarle yo, o llamarle alguien distinto. Según el Criterio Psicológico Amplio de identidad, le llamaríamos alguien distinto. Pero ninguna de estas descripciones incurriría en un error fáctico. Las dos son descripciones del mismo hecho. Si damos una de ellas para favorecer a alguna concepción de lo que importa, nuestra descripción podría ser una mala descripción. Podría conllevar una concepción insostenible de lo que importa. Pero tenemos que decidir lo que importa *antes de* elegir nuestra descripción.

Supongamos que acepto estas afirmaciones. Como reduccionista, ¿debería estar egoístamente preocupado por el futuro de esa persona? ¿Debería estarlo, aunque sé que la continuidad física no puede hacer que sea verdadero, como hecho adicional, que esta persona vaya a ser yo? Al decidir lo que importa, tengo que dejar de lado todos los pensamientos acerca de mi identidad. La cuestión de la identidad es aquí vacía. Tengo que preguntar si, *en sí misma*, la continuidad física justifica la preocupación egoísta.

Creo que la respuesta sería No. Como defendí, los que creen en el Criterio Físico no pueden exigir convincentemente la continuidad de todo el cuerpo. No puede importar el que yo reciba un órga-

no trasplantado, si funciona igual de bien. Todo lo que se podría decir que importa es que una cantidad suficiente de mi cerebro siga existiendo.

¿Por qué se debe resaltar el cerebro de este modo? La respuesta tiene que ser: «Porque el cerebro es el portador de la continuidad psicológica, o relación R». Si por esto se resalta el cerebro, la continuidad del cerebro no importaría cuando *no* fuese el portador de la relación R. La continuidad del cerebro no sería, en ese caso, más importante que la de cualquier otra parte del cuerpo. Y la continuidad de estas otras partes no importa en absoluto. Daría igual que fueran reemplazadas por duplicados suficientemente similares. Diríamos lo mismo del cerebro. La continuidad del cerebro importa si va a ser la causa de que se dé la relación R. Si R *no* se va a dar, la continuidad del cerebro no tendría significación para la persona cuyo cerebro es ahora. No justificaría la preocupación egoísta.

Los reduccionistas no pueden afirmar convincentemente que sólo importe la continuidad física. Como mucho, pueden afirmar que tal continuidad es parte de lo que importa. Como mucho, pueden defender (2), la tesis de que la relación R no importaría si no tuviera su causa normal, parte de la cual es la continuidad física.

Creo que (2) también es insostenible. Pienso que la continuidad física es el elemento menos importante en la existencia continua de una persona. Lo que valoramos, en nosotros y en los demás, no es la existencia continua de los mismos cerebros y los mismos cuerpos particulares. Lo que valoramos son las diversas relaciones entre nosotros y los demás, aquellos a quien amamos y aquello que amamos, nuestras ambiciones, logros, compromisos, emociones, recuerdos y otros diversos rasgos psicológicos. Algunos de nosotros también querríamos que nosotros mismos o los demás siguiéramos teniendo cuerpos muy parecidos a nuestros cuerpos presentes. Pero esto no es querer que sigan existiendo los mismos cuerpos particulares. Pienso que, si más adelante va a haber una persona que estará R-relacionada conmigo como soy ahora, tiene muy poca importancia el que tenga mi cerebro y mi cuerpo actuales. Pienso que lo que fundamentalmente importa es la relación R, aunque no tenga

su causa normal. De manera que daría igual que mi cerebro fuera reemplazado por un duplicado exacto [55].

Si una persona va a estar R-relacionada conmigo, su cuerpo debería ser también lo suficientemente parecido a mi cuerpo actual como para permitir una conexividad psicológica plena. Lo cual no sería el caso, por ejemplo, si fuera del sexo opuesto. Y para unas cuantas personas, las que son muy bellas, por ejemplo, debería también haber igualdad física exacta. Las afirmaciones sobre esta igualdad las omitiré en el futuro.

Aceptar esta concepción puede afectar a nuestras creencias y actitudes sobre nuestra propia vida. Pero la cuestión está clarísima en el caso imaginario del teletransporte. Según mi modo de ver, mi relación con mi Réplica contiene lo que fundamentalmente importa. Esta relación es casi tan buena como la de la supervivencia corriente. Juzgada desde el punto de vista de la Concepción No-Reduccionista, la supervivencia corriente es, según mi modo de ver, un poco mejor que —o casi tan mala como— ser destruido y replicado. Sería irracional, por consiguiente, pagar mucho más por un viaje convencional en nave espacial.

A muchos les daría miedo el teletransporte. Admito que, a cierto nivel, a mí también me lo podría dar. Pero, como he defendido, ese miedo no puede ser racional. Puesto que sé lo que va a ocurrir exactamente, no me puede dar miedo que sea el peor de dos resultados lo que va a ocurrir.

Mi relación con mi Réplica es R sin su causa normal. La anormalidad de la causa me parece trivial. Volvamos a considerar los ojos artificiales que harían recuperar la vista a los que se han quedado ciegos. Supongamos que esos ojos les dieran sensaciones visuales idénticas a las que implica la visión normal, y que estas sensaciones les proporcionaran creencias verdaderas sobre lo que pueden ver. Esto sería con seguridad tan bueno como la visión normal. No sería

[55] Tanto Nagel como Williams se inclinan a pensar que la continuidad física es lo que importa. Pero mientras que Nagel cree que recibir un cerebro duplicado sería tan malo como la muerte, Williams cree que esto sería una mera ampliación trivial de los tipos de cirugía existentes. Véase Williams (2), p. 47.

plausible rechazar estos ojos porque no fueran la causa normal de la visión humana. Habría algunas razones para que no nos gustaran los ojos artificiales, puesto que harían la apariencia de las personas inquietante para los demás. Pero en el teletransporte no hay nada análogo a esto. Mi Réplica, aunque producida artificialmente, va a ser exactamente como yo en todos los aspectos. Tendrá un cerebro y un cuerpo normales.

Probablemente decidiríamos no llamar yo a mi Réplica. Si tomamos esta decisión, deberíamos considerar este caso como el de mi división. Es un caso en el que hay una respuesta mejor a una pregunta vacía. Si estoy a punto de dividirme, lo mejor es decir que ninguna de las personas resultantes va a ser yo. Pero esto no implica que deba considerar la división como similar a la muerte. No estamos decidiendo cuál de varios resultados será lo que ocurra; simplemente estamos eligiendo una de varias descripciones de un único resultado. Puestas así las cosas, nuestra elección de una descripción es irrelevante para la pregunta de cómo debo considerar este resultado.

Lo mismo ocurre en el caso en que seré teletransportado. Mi actitud ante este resultado no debería estar afectada por nuestra decisión de si llamar o no a mi Réplica yo. Conozco todos los hechos aunque no haya tomado aún esa decisión. Si decidimos no llamar a mi Réplica yo, el hecho

(a) de que mi Réplica no vaya a ser yo

nada más que consistiría en el hecho

(b) de que no habrá continuidad física

y

(c) de que, puestas así las cosas, R no tendrá su causa normal.

Como (a) no consistiría en otra cosa que en (b) y (c), la voy a ignorar. Mi actitud dependería de la importancia de los hechos (b) y (c). Estos hechos son todo lo que hay en que mi Réplica no sea yo.

Cuando vemos que esta última afirmación es verdadera, creo que no podemos afirmar racionalmente que (c) importe mucho. No puede importar mucho que la causa sea anormal. Es el *efecto* el que importa. Y este efecto, el darse de la relación R, es en sí el mismo. Es cierto que, si este efecto tiene la causa anormal, lo podemos describir de un modo diferente. Podemos decir que, aunque mi Réplica sea psicológicamente continua conmigo, no será yo. Lo cual, sin embargo, no es una diferencia adicional en lo que ocurre, más allá de la diferencia en la causa. Si decido no apretar el botón y pagar mucho más por un viaje convencional en nave espacial, tengo que admitir que esto es sólo porque no me gusta la idea de un método anormal de causación. No puede ser racional preocuparme mucho por la anormalidad de esta causa.

Comentarios similares se aplican a la existencia continua del cerebro y el cuerpo actuales de alguien. Puede ser racional querer que el cuerpo de mi Réplica sea como mi cuerpo actual. Pero este es el deseo de un cierto tipo de cuerpo, no el deseo del mismo cuerpo particular. ¿Por qué iba a desear yo que *este* cerebro y *este* cuerpo lleguen a Marte? Una vez más, el miedo natural es que sólo esto asegura que *yo* vaya a llegar a Marte. Pero así se asume otra vez que si yo llego o no llego a Marte es, aquí, una pregunta real. Y debemos concluir que esta es una pregunta vacía. Aunque esta pregunta tenga una respuesta que sea la mejor, podemos saber con exactitud lo que va a ocurrir antes de decidir cuál es esa respuesta. Puestas así las cosas, ¿puede preocuparme mucho racionalmente si el cerebro y el cuerpo de esta persona van a ser o no van a ser mi cerebro y mi cuerpo actuales? Creo que, mientras que puede que no sea irracional preocuparse un poco, preocuparse mucho sería irracional.

¿Por qué no sería irracional preocuparse un poco? Se podría tratar de algo parecido al deseo de conservar el mismo anillo de boda en vez de cambiarlo por uno nuevo exactamente igual. Comprendemos el deseo sentimental de conservar el mismo anillo que llevamos en la ceremonia de boda. Del mismo modo, a lo mejor no es irracional tener una ligera preferencia por que la persona en Marte tenga mi cerebro y mi cuerpo actuales.

Queda una pregunta. Si va a haber una persona que estará R-relacionada conmigo, ¿importaría el hecho de que esta relación no tuviera una causa fiable?

Hay una razón obvia para preferir, de antemano, que la causa sea fiable. Supongamos que el teletransporte funcionara perfectamente en unos pocos casos, pero que en la mayoría fuese un completo fracaso. En estos pocos casos, la persona de Marte sería una perfecta Réplica mía. Pero en la mayor parte de los casos sería totalmente distinta de mí. Si los hechos fueran estos, sería claramente racional pagar el billete más caro por un viaje en nave espacial. Pero esto es irrelevante. Preguntaríamos, «En los pocos casos en que mi Réplica va a estar completamente R-relacionada conmigo, ¿importaría que R no tuviese una causa fiable?».

Creo que la respuesta sería de nuevo No. Supongamos que para una enfermedad hay un tratamiento que no es digno de confianza. En la mayoría de los casos no logra nada. Pero en unos pocos obtiene una cura completa. En estos casos, sólo importa el efecto. Este efecto es perfecto, aunque su causa no sea digna de confianza. Diríamos lo mismo de la relación R. Concluyo que, de las respuestas que describí, deberíamos aceptar (4). En nuestra preocupación por nuestro propio futuro, *lo que fundamentalmente importa es la relación R, con cualquier causa.*

97. El caso de la línea secundaria

He defendido que el teletransporte sería más o menos tan bueno como la supervivencia corriente. Otro desafío a esta afirmación viene del caso de la línea secundaria. Supongamos que el nuevo escáner no ha destruido ni mi cerebro ni mi cuerpo, pero que sí ha dañado mi corazón. Estoy aquí en la Tierra, y espero morir dentro de unos días. Usando el Intercom, veo a mi Réplica en Marte y hablo con ella. Me asegura que continuará mi vida dónde yo la deje.

¿Cuál sería mi actitud aquí? Estoy a punto de morir. ¿Contiene mi relación con *esta* Réplica lo que importa? Ella es del todo psicológicamente continua no conmigo como soy ahora sino conmigo

como era esta mañana, cuando apreté el botón verde. ¿Es esta relación tan buena como la supervivencia?

Tal vez sea difícil de creer que lo es. Pero también es difícil de creer que puede importar mucho el que mi vida se solape brevemente con la vida de mi Réplica.

Puede servir de ayuda considerar el caso de

> *La pastilla de dormir*. Ciertas pastillas de dormir reales causan amnesia *retrógrada*. Puede ocurrir que, si tomo una de esas pastillas, seguiré despierto una hora pero después de una noche de sueño no tendré recuerdos de la segunda mitad de esa hora.
>
> He tomado en efecto esas pastillas, y he visto cómo son los resultados. Supongamos que tomé una pastilla de esas hace casi una hora. La persona que se despierte en mi cama mañana no será psicológicamente continua conmigo como soy ahora. Será psicológicamente continua conmigo tal y como yo era hace media hora. Ahora estoy en una *línea secundaria psicológica*, que terminará pronto cuando me quede dormido. Durante esta media hora, soy psicológicamente continuo conmigo mismo en el pasado. Pero no soy ahora psicológicamente continuo conmigo mismo en el futuro. Nunca recordaré después lo que haga, piense o sienta durante esta media hora. Esto significa que, en algunos aspectos, mi relación conmigo mismo mañana es como una relación con otra persona.
>
> Supongamos, por ejemplo, que he estado preocupado por alguna cuestión práctica. Ahora veo la solución. Como está claro lo que debería hacer, me formo una intención firme. En el resto de mi vida sería suficiente con formarme esa intención. Pero cuando estoy en esta línea secundaria psicológica, no es suficiente. Después no recordaré lo que he decidido ahora y no me despertaré con la intención que me he formado ahora. Por eso tengo que comunicarme conmigo mismo mañana como si me estuviera comunicando con alguien distinto. Tengo que escribirme a mí mismo una carta, contando mi decisión y mi nueva intención. Luego tengo que colocarla donde no tenga más remedio que verla mañana.
>
> En efecto, no tengo ningún recuerdo de haber tomado tal decisión ni de haber escrito la carta. Pero una vez encontré una carta así bajo mi navaja de afeitar.

En un aspecto, este caso es como el de la línea secundaria. Y sería nada más que una variante de este caso, en la que, aunque vivo

unos cuantos días después de salir del cubículo, mi Réplica no sería creada hasta después de que yo haya muerto. Pero en el caso que estamos considerando mi vida se solapa con la de mi Réplica. Hablamos por el Intercom. No hay análogo de esto en el caso de la pastilla de dormir.

Sí que puede encontrarse el análogo en mi Examen de Física imaginario. En este caso, divido mi mente durante diez minutos. En mis dos corrientes de conciencia sé que estoy teniendo ahora pensamientos y sensaciones en mi otra corriente. Pero en cada corriente no tengo conciencia de mis pensamientos y sensaciones en mi otra corriente. Mi relación conmigo mismo en mi otra corriente es otra vez como mi relación con otra persona. Tendría que comunicarme de manera pública. Podría en una de las corrientes escribirme una carta a mí mismo en mi otra corriente. Con una mano pondría entonces la carta en mi otra mano.

Esto es como mi situación en el Caso de la Línea Secundaria. Puedo imaginarme teniendo una mente dividida. Puestas así las cosas, no necesito asumir que mi Réplica en Marte sea alguien distinto. Aquí en la Tierra no soy consciente de lo que está ahora pensando mi Réplica en Marte. Esto es como el hecho de que, en cada una de mis corrientes de conciencia en mi Examen de Física, no soy consciente de lo que estoy ahora pensando en mi otra corriente. Puedo creer que ahora sí que tengo otra corriente de conciencia ajena, de la cual, en esta corriente, soy ahora inconsciente. Y, si sirve de ayuda, puedo adoptar esta forma de ver las cosas en relación con mi Réplica. Puedo decir que ahora tengo dos corrientes de conciencia, una aquí en la Tierra y otra en Marte. Esta descripción no puede incurrir en un error fáctico. Cuando le hablo a mi Réplica en Marte, esto es simplemente como la comunicación en el Examen de Física conmigo mismo en mis dos corrientes.

El caso real de la pastilla de dormir proporciona una estrecha analogía con uno de los rasgos especiales del caso de la línea secundaria: el hecho de que estoy en una línea secundaria psicológica. El Examen de Física imaginario proporciona una estrecha analogía con el otro rasgo especial: que mi vida se solapa con la de mi Réplica. Cuando consideramos estas analogías, esto parece suficiente para

defender la afirmación de que, cuando estoy en la Línea Secundaria, mi relación con mi Réplica contiene casi todo lo que importa. Tal vez resulte algo inconveniente que mi Réplica vaya a ser psicológicamente continua no conmigo mismo tal y como soy ahora, sino conmigo mismo como era esta mañana cuando apreté el botón verde. Pero estas relaciones son sustancialmente las mismas. Que mi vida se solape brevemente con la de mi Réplica marca sólo una pequeña diferencia.

Si el solapamiento fuese largo, esto *sí* que significaría una diferencia. Supongamos que soy un anciano que está a punto de morir. Me sobrevivirá alguien que una vez fue una Réplica mía. Cuando esta persona empezó a existir hace cuarenta años, era psicológicamente continua conmigo tal y como yo era entonces. Desde aquel momento vivió su propia vida durante cuarenta años. Estoy de acuerdo con que mi relación con *esta* Réplica, aunque mejor que la muerte corriente, no es ni con mucho tan buena como la supervivencia corriente. Pero esta relación sería casi tan buena si mi Réplica fuese psicológicamente continua conmigo como yo era hace diez días o diez minutos. Como sostiene Nozick, los solapamientos tan breves como este no puede pensarse racionalmente que tengan mucha significación [56].

Aunque mis dos analogías parecen bastar para la defensa de esta afirmación, admito que este es uno de los casos en que mi modo de ver las cosas es más difícil de creer. *Antes* de apretar el botón verde, puedo creer más fácilmente que mi relación con mi Réplica contiene lo que fundamentalmente importa en la supervivencia corriente. Puedo mirar hacia el futuro por la Línea Principal en la que quedan cuarenta años de vida. *Después* de haber apretado el botón verde, y de haber hablado con mi Réplica, no puedo mirar hacia el futuro de la misma manera por la Línea Principal. Mi preocupación por el futuro necesita ser redirigida. Tengo que tratar de dirigir esta preocupación hacia atrás por la Línea Secundaria hasta más allá del punto de división, y luego hacia delante por la Línea Principal. Esta maniobra psicológica sería difícil. Pero esto no es sorprendente. Y

[56] *Op. cit.*, p. 44.

como no es sorprendente, esta dificultad no supone un argumento suficiente contra lo que he afirmado sobre este caso.

98. Personas-serie

He negado que la identidad personal sea lo que importa. Según mi concepción, lo que fundamentalmente importa, en nuestra preocupación por nuestro propio futuro, es el darse de la relación R, con cualquier causa. Esto sería lo que importa, aunque no coincidiera con la identidad personal.

Según el modo de ver las cosas de Nagel, lo que yo soy esencialmente es mi cerebro. Y lo que fundamentalmente importa es la existencia continua de este cerebro. Pienso que doy respuesta a los argumentos de Nagel en el Apéndice D. Pero estos argumentos quizás no convenzan. Por eso vale la pena explicar cómo podrían ser ambas verdaderas, la concepción de Nagel y una forma revisada de la mía.

Supongamos que la idea de Nagel es verdadera. Lo que fundamentalmente importa, para mí, es la existencia continua de mi cerebro. Como según la concepción de Nagel yo soy esencialmente mi cerebro, no puedo decidir adoptar una idea diferente sobre mí mismo. Pero puedo hacer otra cosa.

Nagel describe el concepto de *persona-serie*. Mientras que una persona es esencialmente, según la opinión de Nagel, un cerebro particular incorporado, una persona-serie sería potencialmente una serie R-relacionada de cerebros incorporados. Ahora no podemos solucionar el problema de que nuestros cuerpos envejezcan y se deterioren. Nagel imagina una comunidad en la que la tecnología aporta una solución. En ella, desde que cumplen 30 años, todos entran una vez por año en un replicador escáner. Esta máquina destruye el cerebro y el cuerpo de la persona, produciendo una Réplica que está R-relacionada con ella, y que tiene un cuerpo exactamente igual, sólo que ni envejece ni se deteriora. Nagel afirma que, para las personas-serie de esta comunidad, no sería irracional utilizar este replicador escáner. Según su criterio de identidad, cada una de estas personas-serie seguiría existiendo, trasladándose a un nuevo cerebro y a un nuevo cuerpo cada año. Cada persona-serie siempre tendría un cerebro y un cuerpo con la juventud, la apariencia y el vigor de su cerebro y su cuerpo a la edad de 30 años.

Pero estas personas-serie podrían tener accidentes mortales. Por eso añado yo el detalle de que, como precaución, cada una se hace su cianotipo cada día. Con este añadido, estas personas-serie son potencialmente inmortales, pudiendo disfrutar de la eterna juventud. Según la mayor parte de las teorías que se aceptan ahora, el universo o bien se expanderá indefinidamente, o bien se colapsará, en una inversión del «Big Bang». La mayoría de los físicos dan por sentado que, en cualquiera de las dos alternativas, todas las formas de vida se harán imposibles. Si esto no fuera así, estas personas-serie podrían vivir para siempre [57].

Asumamos que acepto la concepción de Nagel. Yo soy esencialmente mi cerebro, y lo que fundamentalmente importa, en mi preocupación por mi futuro, es la existencia continua de este cerebro. Si acepto este modo de ver las cosas, tengo que haber sido empujado de mala gana a esta conclusión. Prefería, con mucho, mi vieja concepción. Aunque no pueda cambiar mi parecer sobre lo que soy, puedo hacer lo que sigue. Ahora estás leyendo frases que mecanografié en noviembre de 1982. Esta frase te dice que, en el resto de este libro, *los pronombres se usan para referir a personas-serie*. Si la concepción de Nagel es falsa, puede que esto no cambie lo que los pronombres significan. Cada persona puede ser una persona-serie. Esto sería así si nuestro criterio de identidad fuera la relación R con cualquier causa, puesto que este es también el criterio de identidad de las personas-serie.

Si la opinión de Nagel es cierta, el resto de este libro usa los pronombres en un nuevo sentido. Refieren no a personas sino a per-

[57] En un *tour de force*, Freeman Dyson afirma que la vida podría continuar para siempre incluso en un universo en expansión indefinida. El argumento exige la asunción de que nosotros podríamos emigrar de cuerpos orgánicos a cuerpos inorgánicos. Véase «Time Without End: Physics and Biology in an Open Universe» [«Tiempo sin fin: Física y Biología en un universo abierto»], *Review of Modern Physics*, vol. 51 (1979), p. 447.

sonas-serie. Por ejemplo, las palabras, «yo», «me» y «mí» no refieren a la persona, Derek Parfit. Refieren a la persona-serie cuyo cerebro y cuerpo actuales son también el cerebro y el cuerpo de Derek Parfit. Como las palabras «yo», «me» y «mí» se usan en este sentido nuevo, su viejo sentido se expresa con las palabras «viejo-yo», «viejo-me» y «viejo-mí». Observaciones similares se aplican a los otros pronombres.

Según la idea de Nagel, ¿cuál es la relación entre viejo-yo y yo la persona-serie? Puede servir de ayuda recordar a un ser mítico: el fénix. Según el criterio de identidad de aves, un ave deja de existir si arde hasta que se convierte en cenizas. Si un fénix existiese, no sería un ave particular. Sería una serie de aves, o una *ave-serie*. Un fénix tendría en cualquier momento el cuerpo de un ave particular. Pero cuando este ave arde hasta convertirse en cenizas, sólo el ave deja de existir. El fénix viene otra vez a la vida en el cuerpo de un ave nueva, que resurge de las cenizas. Como una persona-serie particular, un fénix particular tendría, de esta forma, una serie de cuerpos diferentes.

No ha existido nunca ningún fénix. Pero hay muchas personas-serie. Estas frases están siendo mecanografiadas por una persona-serie, yo. También están siendo mecanografiadas por una persona: viejo-yo. Esta persona se llama Derek Parfit. Yo la persona-serie, por la presente, me llamo a mí mismo *Fénix Parfit*. Como mi cuerpo actual es también el cuerpo de Derek Parfit, los dos estamos mecanografiando estas frases. Y los dos estamos teniendo los mismos pensamientos y las mismas experiencias. Pero, aunque ahora estemos relacionados de esta manera extremadamente íntima, *si* la idea de Nagel es verdadera, somos dos individuos diferentes. La diferencia entre nosotros es esta: según la opinión de Nagel, si viejo-yo fuera teletransportado, eso mataría a viejo-yo, la persona. Pero no me mataría a mí, la persona-serie. Esta diferencia es suficiente como para hacer de viejo-yo y yo individuos diferentes.

Puedes dudar de que yo, esta persona-serie, realmente exista. Puedes pensar que no puede hacerse existir a los entes simplemente inventando nuevos conceptos. Es verdad. No me hizo existir ni

la invención de Nagel del concepto de persona-serie, ni el mecanografiado de las frases de arriba que pueden haber dado nuevos significados a «yo», «me» y «mí». Dado lo que significa el concepto de «persona-serie», yo la persona-serie comencé a existir cuando viejo-yo la persona comenzó a existir. Y es muy probable que ambos dejaremos de existir al mismo tiempo. Esto es muy probable porque es extremadamente improbable que, en el transcurso de la vida de viejo-yo, el teletransporte llegue a ser posible.

Las lenguas evolucionan. Cuando inventamos un concepto nuevo, puede que descubramos que se aplica a partes de la realidad. El concepto de fénix no tiene aplicación a nada. Pero el concepto de persona-serie se aplica las mismas veces que el de persona. Sea o no sea cierta la concepción de Nagel, ahora existen varios miles de millones de personas-serie. Si la concepción de Nagel es cierta, por cada persona que existe hay una persona-serie que existe y que está *muy* íntimamente relacionada con esa persona. Dada esta relación extremadamente íntima, la distinción entre estos individuos no vale la pena establecerla prácticamente nunca. Vale la pena establecerla sólo en este punto de mi discusión de lo que importa. Estoy suponiendo que la concepción de Nagel es verdadera. Estoy suponiendo que lo que fundamentalmente importa para viejo-yo es la existencia continua de mi cerebro presente. Aunque creamos en esto, todavía podemos pensar que la existencia continua de nuestros cerebros presentes *no* es lo que importa. Podemos afirmar que lo que importa es la relación R. Esto es lo que importa *para nosotros, las personas-serie.*

Con un concepto nuevo a veces podemos dar una mejor descripción de la realidad. Esto ocurrió con los conceptos de átomo y molécula. Esta clase de mejora está clara. Si la idea de Nagel es correcta, el concepto de persona-serie también hace posible una mejor descripción de la realidad. Pero esta mejora no está tan clara. El concepto de persona-serie no sólo permite a los mismos seres dar una mejor descripción de la realidad. Permite a seres *diferentes* tanto proclamar su existencia cuanto dar esta mejor descripción.

Nagel menciona un tercer concepto, el de *persona-día*. Tal existencia de una persona conlleva necesariamente una corriente ininte-

rrumpida de conciencia. Para las personas-día, el sueño es la muerte. Cuando consideramos este concepto, de nuevo descubrimos que tiene aplicación a la realidad. En un momento dado hay tantas personas-día viviendo como personas conscientes viviendo. Pero durante un año entero el número de personas-día que han vivido excederá enormemente el número de personas vivientes.

El concepto de persona-día es peor que el concepto de persona, porque lo que importa no es que la corriente de conciencia no se interrumpa. Lo que importa es la relación R. Aunque tiene aplicación a la realidad, el concepto de persona-día selecciona partes de la realidad con límites que no tienen importancia. Creemos, de manera plausible, que no importa que haya interrupciones en una corriente de conciencia, porque no destruyen la continuidad psicológica.

Si la concepción de Nagel es verdadera, lo que le importa a una persona es la existencia continua de un cerebro particular. Pero si preguntamos por lo que es importante respecto de nosotros mismos, y de nuestras vidas y nuestras relaciones con los demás, la existencia continua de cerebros particulares no parece ser lo que importa. Más bien, como ya he dicho, lo que es más importante es la relación R, conexividad y/o continuidad psicológica, con cualquier causa. Si esto es así, y la opinión de Nagel es correcta, el concepto de persona es peor que el concepto de persona-serie, y peor de un modo parecido. Según la opinión de Nagel, lo que importa para una persona es la existencia continua de su cerebro presente. Lo que importa para una persona-serie es la relación R, con cualquier causa. El concepto de persona-serie es mejor porque apela a lo que es más importante.

Una persona no puede negar que es una persona. Y, si la concepción de Nagel es verdadera, una persona no puede volverse una persona-serie. Pero una persona-serie puede empezar a hablar por la boca que las dos comparten. La persona-serie puede proclamar su existencia, ponerse un nombre, y afirmar que todos los pronombres que escribe o pronuncia se referirán a personas-serie. Todas las demás personas-serie podrían hacer lo mismo. Todas las vidas humanas futuras serían vividas entonces por seres que se

consideraríian a sí mismos personas-serie. Estas vidas también serían vividas por personas. Pero las personas tendrían ahora un papel subordinado, puesto que rara vez se referirían a los sí mismos antiguos.

Si la concepción de Nagel no es correcta, los sucesos que acabo de describir puede que no presenten ninguna diferencia. Puede que nuestras creencias sobre el criterio de identidad no sean capaces de incluir unos cuantos casos reales, como los de las personas con los hemisferios divididos. Y está claro que estas creencias no son capaces de incluir muchos casos imaginarios. Como las personas no son entidades que existan separadamente, distintas de sus cerebros y de sus cuerpos y de sus experiencias, las preguntas por la identidad personal son, en estos casos imaginarios, vacías. En estos casos podemos *dar* respuesta a estas preguntas, ampliando con ello nuestras creencias sobre el criterio de identidad personal. Una ampliación posible haría de este criterio el darse no ramificado de la relación R, con cualquier causa. Según este criterio, las personas *son* personas-serie. La distinción trazada arriba desaparece.

Si la concepción de Nagel es correcta, no puedo hacer estas afirmaciones. Según su modo de pensar, cada persona es esencialmente su cerebro, y lo que fundamentalmente importa, para cada persona, es la existencia continua de ese cerebro. Puestas así las cosas, las personas no son personas-serie. Los sucesos que describí arriba *sí* que supondrían una diferencia. Si todas las personas-serie proclamaran su existencia, y empezaran a usar pronombres para referirse a sí mismas, esto sería una mejora. Cada persona-serie está relacionada muy estrechamente con una persona particular. Sería mejor que, en cada uno de estos pares, la persona-serie asumiera el papel conductor. El concepto de persona-serie selecciona partes de la realidad de un modo menos arbitrario. Nosotros las personas-serie podemos negar que lo que importa sea, para nosotros, la existencia continua de nuestros cerebros. Podemos afirmar que, como he sostenido, lo que fundamentalmente importa es la relación R, con cualquier causa.

99. ¿SOY UNA MUESTRA O UN TIPO?

Williams considera un caso en el que una persona tendría muchas réplicas coexistentes. Y sugiere una nueva descripción de un caso como este. Describe el concepto de una *persona-tipo*. Supongamos que hay una persona particular, Mary Smith. Y supongamos que el replicador escáner produce muchas réplicas de Mary Smith, tal y como ella es en un momento dado. Todas estas réplicas serán Marys Smith. Serán diferentes *muestras*, o instancias, de la misma persona-tipo. Si un caso así ocurriese, habría varias preguntas sobre lo que importa. Asumamos que en el caso interviene el viejo escáner, que destruye el cerebro y el cuerpo originales de Mary Smith. Antes de apretar el botón verde, ¿qué debería pensar Mary Smith de su relación con sus réplicas futuras? ¿Lo que importa es que su cerebro actual siga existiendo? ¿O será casi tan bueno que vaya a haber después muestras vivas de su tipo?

Esta es la pregunta por lo que debe interesar a la Mary Smith original. Williams no la discute, sino que escribe de una manera fascinante acerca de otra pregunta:

«Como no estamos suponiendo que las personas-muestra, impresas a partir del prototipo, tengan experiencias que se comuniquen entre sí... serán afectadas de una manera divergente por diferentes experiencias, y tenderán a hacerse cada vez más distintas. Contempladas como copias del prototipo, se convertirán en copias cada vez más borrosas o sobrescritas; contempladas en sí mismas, se convertirán en personalidades cada vez más individuales. Lo cual podría ser bienvenido, porque alguien que amara a una de estas personas-muestra podría muy bien amarla no porque fuese una Mary Smith, sino a pesar del hecho de que fuese una Mary Smith... Cuanto más se diferenciaran las *Mary Smith*, tanto más segura la influencia que el amante podría sentir que tenía sobre lo que amaba particularmente.

Si alguien amara a una persona-muestra igual que Mary Smith, entonces podría muy bien no estar claro que la persona-muestra era realmente lo que amaba. A quien ama es a *Mary Smith*, y esto es amar a la persona-tipo. Podemos ver vagamente a qué se parecería esto. Se parecería a amar una obra de arte en algún medio reproducible.

Uno podría empezar a comparar, como si dijésemos, representaciones del tipo; y querer estar cerca de la persona que uno amase sería como querer escuchar vivamente una representación de *Fígaro*, incluso una mediocre —igual que uno irá a la representación pueblerina improvisada de *Fígaro* antes que no escuchar *Fígaro* en absoluto—. De este modo, uno iría a ver a la muy gastada Mary Smith que está en su localidad antes que no ver a ninguna Mary Smith en absoluto.

Mucho de lo que llamamos querer una persona empezaría a quebrarse bajo estas condiciones, y el reflexionar sobre ello puede animarnos a no menospreciar la situación profundamente basada en el cuerpo que realmente tenemos. Mientras que en el actual estado de cosas amar a una persona no es exactamente lo mismo que amar a un cuerpo, quizás decir que son básicamente lo mismo sea más tremendamente engañoso antes que un profundo error metafísico; y si no suena muy noble, las alternativas que surgen tan enérgicamente de suspender la situación presente tampoco suenan muy espirituales» [58].

¿Amar a una persona es básicamente lo mismo que amar el cuerpo de esa persona? Williams admite que esta afirmación es «tremendamente engañosa»; pero sugiere que es mejor que cualquier alternativa. Y nos advierte de que no menospreciemos «la situación profundamente basada en el cuerpo que realmente tenemos». Una situación diferente, en que las personas tuvieran muchas réplicas coexistentes, pondría en peligro mucho de lo que valoramos.

Creo que deberíamos aceptar esta última afirmación. Pero no deberíamos afirmar que amar a una persona es básicamente lo mismo que amar el cuerpo de esa persona. *Hay* una alternativa mejor.

Williams puede haber razonado como sigue. A no ser que lo que amo sea un cuerpo particular, no puedo amar a un individuo. Supongamos que amo a la Mary Smith original. Una máquina destruye su cerebro y su cuerpo y produce una Réplica. Si se transfie-

[58] Williams (2), pp. 80-1. Véase también Brennan (1) y (2).

re mi amor por Mary Smith a su Réplica, esto sugiere que lo que amo no es un individuo sino una persona-tipo. Y si consideramos lo que conlleva un amor así, lo que encontramos es inquietante.

Estoy de acuerdo en que un amor así sería muy diferente, e inquietante. Pero rechazo el razonamiento que se acaba de dar.

Deberíamos considerar *dos* clases de casos imaginarios. Uno es la comunidad que Nagel imaginó. En esta comunidad, aunque haya replicación, la relación R nunca toma una forma ramificada. A partir de los 30, Mary Smith utiliza una vez al año el replicador que preserva la juventud, como por otro lado hacen muchos. Si esas máquinas existieran sería *posible* producir varias réplicas coexistentes de un mismo individuo. Pero podemos suponer que, en esta comunidad imaginaria, los individuos han tomado la decisión de no realizar esta posibilidad. Por las razones que detallé en la Sección 90, y que expresa mejor Williams, estos individuos son de la opinión de que la división no es tan buena como la supervivencia ordinaria.

Supongamos que soy una persona que se ha ido a vivir a esta comunidad diferente. Me enamoro de Mary Smith. ¿Cómo debería reaccionar después de que ella haya usado por primera vez el replicador? Afirmo que no sólo amaría de hecho sino que *debo* amar a su Réplica. Y no se trata del «debo» de la moral. Según la mejor concepción de la mejor clase de amor, yo debo amar a este individuo. Es completamente continua desde el punto de vista psicológico con la Mary Smith que yo amaba, y tiene un cuerpo exactamente igual. Si no amo a la Réplica de Mary Smith, sólo podría ser así por una de varias razones, todas malas.

Una razón podría ser que creo en la Concepción No-Reduccionista. Creo que la identidad personal es un hecho adicional profundo, que no se produciría por replicación. No amo a la Réplica de Mary Smith porque creo que ella no es Mary Smith de este modo profundo. Esta reacción carece de justificación puesto que no hay semejante hecho adicional.

Supongamos a continuación que acepto la Concepción Reduccionista, pero creo que la replicación es casi tan mala como la muerte corriente. Cuando Mary Smith aprieta el botón mi reacción viene

a ser de pena. Tal vez más adelante pueda llegar a querer a la Réplica de Mary Smith. Pero dada mi creencia en la maldad de lo que le ocurrió, mi amor no puede ser simplemente transferido sin pena.

Como he sostenido, esta reacción tampoco está justificada. Según la Concepción Reduccionista, deberíamos considerar la replicación casi tan buena como la supervivencia corriente. Como Mary Smith eligió ser replicada podemos dar por sentado que esta era también su opinión. Según esta opinión, mi amor debería transferirse a su Réplica. Además, en esta comunidad imaginaria los individuos son personas-serie. Por tanto, la Réplica de Mary Smith *es* Mary Smith.

Si mi amor no es transferido, podría haber dos explicaciones adicionales. Williams sugiere que amar a una persona es, básicamente, amar a un cuerpo particular. Pero esta clase de amor, o deseo, es como mucho extremadamente rara. Lo que es más común es una obsesión puramente física o sexual por el cuerpo de una persona, una obsesión que no se interesa en la psicología de esa persona. Pero esto *no* es amor de un cuerpo particular. Como escribe Quinton, en el caso de tales obsesiones, «no se requiere ningún cuerpo humano concreto, sólo uno de una clase más o menos precisamente demarcada» [59]. Supongamos que yo estuve físicamente obsesionado

[59] Quinton, en Perry (1), p. 66. Aunque es materialista, Quinton piensa que lo que importa es la Relación R, y sustituye mi aséptica etiqueta por «el alma». Y continúa: «donde la preocupación por el alma está completamente ausente, no hay ningún interés en la identidad individual, sólo interés en la identidad de tipo». Y escribe:

> «En nuestras relaciones generales con otros seres humanos, sus cuerpos carecen por lo general de importancia intrínseca. Los usamos como convenientes dispositivos de reconocimiento que nos permiten localizar sin dificultad el carácter persistente y los complejos de memoria en los que estamos interesados, que amamos o que nos gustan. Sería terrible que un complejo con el que estamos implicados emocionalmente pasara a tener una apariencia física monstruosa o repulsiva. Sería socialmente embarazoso que se mantuviese pasando de cuerpo en cuerpo mientras la mayoría de estos complejos se quedaran sin moverse, y sería desconcertante y pesado que semejante cambiar de un lado a otro se generalizara, porque sería un negocio laborioso encontrar dónde están los amigos y los parientes de uno. Pero que nuestra preocupación y nuestro cariño seguirían al complejo de carácter y memoria, y no a su asociado corporal original, es con seguridad evidente. En el caso de un andar cambiando genera-

con el cuerpo de Mary Smith. Esta obsesión se transferiría a la Réplica de Mary Smith. Sería como un caso en el que el cuerpo con el que estoy obsesionado es el de una gemela. Si esta gemela muriese, mi obsesión podría transferirse al cuerpo de la otra.

El amor corriente no podría transferirse así. Ese amor está interesado en la psicología de la persona amada, y en la vida mental continuamente cambiante de esa persona. Y amar a alguien es un proceso y no un estado fijo. El amor mutuo implica una historia compartida. Por eso, si he amado a Mary Smith durante muchos meses o muchos años, su gemela no puede ocupar su lugar así por las buenas. Pero las cosas son muy diferentes con su Réplica. Si yo he amado a Mary Smith durante meses o años, su Réplica tendrá cuasi-recuerdos completos de nuestra historia compartida.

He afirmado que, si no amo a la Réplica de Mary Smith, es improbable que la explicación sea que yo amaba su cuerpo. Es dudoso que alguien tenga un amor o deseo de ese tipo. La explicación que queda es que mi amor se ha extinguido por ninguna razón en particular. Pero ninguna razón es una mala razón. El amor puede extinguirse así, pero entonces sólo es una clase inferior de amor.

He discutido la alternativa imaginaria de Nagel al mundo real. En ella, las personas son replicadas a menudo, pero nunca hay dos réplicas coexistentes de una persona. La relación R nunca adopta una forma ramificada. Afirmo que, en este mundo, el amor por una persona debería transferirse directamente a su Réplica. Debería transferirse porque la relación de la persona con su Réplica contiene lo que fundamentalmente importa en la supervivencia ordinaria.

Williams sugiere que, en un mundo con replicación, deberíamos distinguir entre personas-tipo y muestras de esos tipos. Pero en el mundo recién descrito, donde la relación R siempre adopta la forma uno-uno, no sería una distinción útil. Describiríamos erróneamente lo que ocurre si dijéramos que cada nueva Réplica es otra muestra de una persona-tipo. Esta descripción ignoraría lo que es de la mayor importancia, la continuidad psicológica y el desarrollo de una vida. Podríamos describir mejor lo que ocurre de una de las dos maneras siguientes.

lizado, estaríamos en la posición de la gente que trata de encontrar a sus íntimos en la oscuridad. Si los cambios fueran tanto frecuentes como espacialmente radicales no dudaríamos en abandonar el intento de identificar personas individuales, con lo que cambiaría todo el carácter de las relaciones entre las personas, y la vida humana sería como una secuencia sin fin de viajes oceánicos más bien cortos.»

Si la idea de Nagel es falsa, nuestro criterio de identidad personal podría ampliarse hasta ser el darse no ramificado de la relación R. Cada individuo en la comunidad imaginaria de Nagel sería entonces una persona. Personas que se trasladarían a cuerpos nuevos una vez al año. Como Mary Smith tendría esos cuerpos nuevos, mi amor por ella debería transferirse directamente. Aunque estas personas cambiarían de cuerpo con frecuencia, el amor por personas particulares no se vería amenazado.

Si la idea de Nagel es verdadera, los individuos de esa comunidad no son personas. Son personas-serie, y eso es lo que ellos creen que son. La persona-serie Mary Smith se traslada a un cuerpo nuevo cada año. Como antes, no está amenazado el tipo de amor que valoramos. Si yo amo a una persona-serie, no estoy amando a una persona-tipo. Estoy amando a un individuo particular, que tiene una historia continua.

Consideremos a continuación la otra alternativa al mundo real: la que Williams imaginó. En este mundo hay muchas réplicas coexistentes de una única persona. La distinción propuesta por Williams sería aquí de utilidad. Consideremos cincuenta réplicas de Greta Garbo tal y como era a los 30 años. Se podrían describir bien diciendo que son diferentes muestras de una persona-tipo. Como Williams afirma, si el objeto de amor es la persona-tipo, esto es muy diferente del amor corriente. No se trataría de la clase de amor que da gran importancia a una historia compartida.

Si yo viviera en un mundo semejante, y fuese una de un conjunto de réplicas, podría considerarme a mí mismo como una muestra de un tipo. ¿Podría en vez de eso considerarme a mí mismo como *el tipo*? Esto sería un cambio radical. En un sentido de la palabra «tipo», si yo fuera una persona-tipo, no podría dejar de existir en absoluto. Aunque no haya ahora muestras de mi persona-tipo,

todavía habría esta persona-tipo. Una persona-tipo sobreviviría incluso a la destrucción del universo, puesto que, en este sentido, un tipo es una entidad abstracta, como un número. No podríamos de ninguna manera considerarnos a nosotros mismos entidades abstractas.

En cualquier sentido de «tipo», habría una gran diferencia entre el amor corriente y el amor a una persona-tipo. La última clase de amor no puede ser mutua. Yo puedo amar a una persona-tipo, pero esta persona-tipo no puede amarme a mí. Un tipo no puede amar más de lo que el número nueve puede amar. Yo no puedo ser amado por la Belleza Típicamente Inglesa ni por la Nueva Mujer Americana. Lo que podría ser cierto es que el amor por mí fuese uno de los rasgos de alguna persona-tipo. Entonces, todas las muchas muestras de este tipo me amarían.

Si esto fuese cierto, podría darse amor mutuo entre una de estas personas-muestra y yo. Podría incluso darse amor mutuo entre dos o tres de estas personas y yo. Pero, como dice Williams, este amor me apartaría de amar a la persona-tipo. Me apartaría de ello, a causa de la importancia creciente de las historias compartidas.

Volvamos ahora a las principales afirmaciones de Williams. Él sugiere que, aunque sea engañosa, hay una profunda verdad en la afirmación de que amar a una persona es amar a un cuerpo particular. Y sugiere además que si el objeto de nuestro amor no fuera un cuerpo particular, amaríamos a una persona-tipo. Semejante amor sería muy diferente del amor corriente, y sería inquietante: amenazaría mucho de lo que valoramos.

Acepto esta última afirmación, pero niego las otras. Siguiendo a Quinton, dudo que alguien ame a un cuerpo particular. Una obsesión puramente física es una obsesión por una clase de cuerpo, o un cuerpo-tipo. Como tal, esto tendría los rasgos inquietantes del amor a una persona-tipo. Tras la muerte de un gemelo, esta obsesión podría transferirse, sin ninguna pena, al cuerpo del otro.

También he negado que, si el objeto de nuestro amor no es un cuerpo particular, tengamos que amar a una persona-tipo. Esto está mejor visto en la alternativa imaginaria de Nagel al mundo real, en la que las personas son replicadas a menudo, pero sólo en la forma uno-uno. En este mundo la relación R traza líneas a través de muchos cuerpos diferentes, pero nunca adopta una forma ramificada. Afirmo que, en semejante mundo, el amor corriente sobreviviría sin cambio. Si la concepción de Nagel fuese falsa, las personas de esta sociedad se desplazarían a cuerpos nuevos cada año, pero todavía serían personas particulares. Si la concepción de Nagel fuese verdadera, serían personas-serie las que se desplazarían a cuerpos nuevos. Pero el amor seguiría siendo amor a un individuo particular. Una persona-serie es un individuo.

Si estas afirmaciones son correctas, puedo mantener una vez más la concepción que he defendido. Lo que importa no es la existencia continua de un cuerpo particular, sino la relación R con cualquier causa.

100. Supervivencia parcial

Antes de considerar vidas reales, echaré un vistazo a una última ráfaga de casos imaginarios. Uno es el contrario de la división: la fusión. La identidad es lógicamente uno-uno, y todo-o-nada. Igual que la división muestra que lo que importa en la supervivencia no es necesario que adopte una forma uno-uno, la fusión muestra que puede tener grados.

Había fusión en los casos centrales del Espectro Combinado. En estos casos, la persona resultante estaría psicológicamente conectada, y aproximadamente en el mismo grado, tanto a mí como a alguien distinto.

Podemos imaginar un mundo en que la fusión fuese un proceso natural. Dos personas se juntan. Mientras están inconscientes, sus dos cuerpos se vuelven uno. Luego se despierta una persona.

Esta persona única podría cuasi-recordar vivir las dos vidas de las dos personas originales. No tendría que perderse ningún cuasi-recuerdo. Pero algunas cosas sí que se tienen que perder. Cualesquiera dos personas que se fusionen juntas tendrían diferentes características, diferentes deseos, intenciones diferentes. ¿Cómo se podrían combinar todas estas cosas?

La respuesta podría ser esta. Algunos de estos rasgos serán compatibles. Estos coexistirían en la única persona resultante. Y algunos serán incompatibles. Éstos, si tienen la misma fuerza, se neutralizarían, y si tienen fuerzas diferentes, los más fuertes se harían más débiles. Estos efectos podrían ser tan predecibles como las leyes que gobiernan los genes dominantes y recesivos.

Aquí van algunos ejemplos. Yo admiro a Paladio, y tengo la intención de visitar Venecia. Estoy a punto de fusionarme con uno que admira a Giotto, y tiene la intención de visitar Padua. La única persona resultante tendrá los dos gustos y las dos intenciones. Y como Padua está cerca de Venecia, las dos pueden realizarse fácilmente. Supongamos a continuación que me encanta Wagner y siempre voto a los socialistas. La otra persona aborrece a Wagner, y vota siempre a los conservadores. La única persona resultante será un votante indeciso sin oído musical.

Como la división, la fusión no encaja en la lógica de la identidad. La única persona resultante no puede decirse que sea la misma persona que cada una de las dos personas originales. La mejor descripción es que la persona resultante no sería ninguna de las dos. Pero, como enseña el caso de la división, esta descripción no implica que estas dos personas deban considerar la fusión como equivalente a la muerte.

¿Cuál debería ser su actitud? Si nosotros estuviéramos a punto de sufrir una fusión de esta clase, algunos de nosotros la podrían considerar como equivalente a la muerte. Lo cual es menos absurdo que considerar la división como equivalente a la muerte. Cuando me divido, las dos personas resultantes serán exactamente como yo. Cuando me fusiono, la única persona resultante no será del todo similar. Esto hace más fácil pensar, cuando nos enfrentamos a la fusión, «No sobreviviré», con lo que seguimos considerando la supervivencia como siendo todo-o-nada.

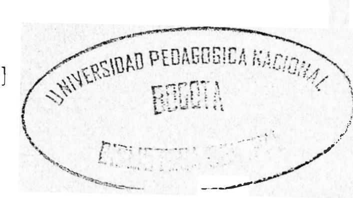
Como he defendido, no hay ningún hecho implicado que sea todo-o-nada. Las dos clases de conexividad, física y psicológica, podrían darse en cualquier grado. ¿Cómo debería considerar un caso en que estas relaciones se dan en grados reducidos?

Podría decirse: «Supongamos que entre una persona resultante y yo hubiera aproximadamente la mitad de la cantidad corriente de estas dos relaciones. Esto sería casi la mitad de bueno que la supervivencia corriente. Si hubiera nueve décimos de estas cantidades corrientes, esto sería aproximadamente nueve décimos tan bueno».

Esta idea es demasiado tosca. Al juzgar el valor que tiene para mí un caso particular de fusión, tenemos que saber lo íntima que es mi relación con la persona resultante. Tenemos que saber también si esta persona tendrá rasgos que considero buenos o malos. La idea recién descrita ignora erróneamente esta segunda cuestión.

Propongo la siguiente concepción. El valor para mí de mi relación con una persona resultante depende no sólo (1) de mi grado de conexividad con esa persona, sino además (2) del valor, según mi opinión, de los rasgos físicos y psicológicos de la misma. Supongamos que la hipnosis me hace perder cinco rasgos no deseados: mi falta de orden, mi pereza, mi miedo a volar en avión, mi adicción a la nicotina, y todos mis recuerdos de mi desdichada existencia. Aquí hay mucho menos que conexividad psicológica completa, pero esto es más que compensado por la eliminación de rasgos malos.

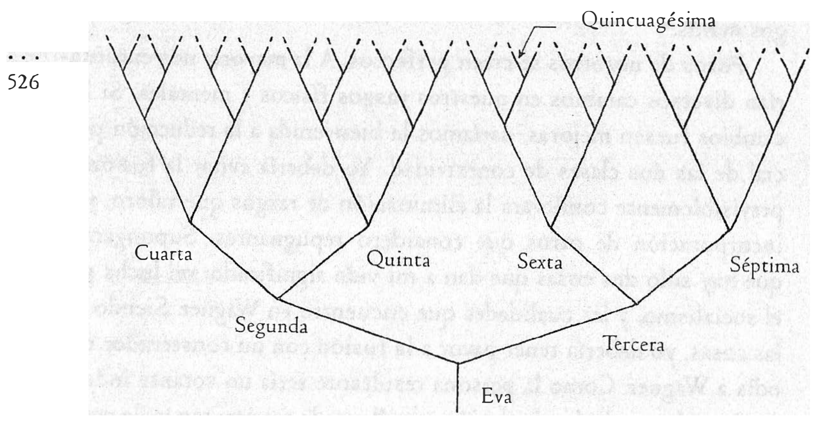
Pocos de nosotros se creen perfectos. A la mayoría nos encantarían diversos cambios en nuestros rasgos físicos y mentales. Si los cambios fuesen mejoras, daríamos la bienvenida a la reducción parcial de las dos clases de conexividad. Yo debería evitar la fusión si previsiblemente conllevara la eliminación de rasgos que valoro, y la incorporación de otros que considero repugnantes. Supongamos que hay sólo dos cosas que dan a mi vida significado: mi lucha por el socialismo, y las cualidades que encuentro en Wagner. Siendo así las cosas, yo debería tener pavor a la fusión con un conservador que odia a Wagner. Como la persona resultante sería un votante indeciso sin oído musical, mi relación con él puede ser casi tan mala como la muerte. Pero otro caso de fusión, aunque implicara un cambio tan grande, yo lo podría considerar mejor que la supervivencia corriente. Podría considerar esos cambios como mejoras. Se podría tratar de añadir un rasgo que me encanta o de eliminar otro que lamento. Las fusiones, como los matrimonios, pueden ser grandes éxitos, o desastres.



Consideremos a renglón seguido algunas personas imaginarias más. Unas que son como nosotros, excepto en lo que respecta a su método de reproducción. Como las amebas, se reproducen por un proceso de división natural. Las vidas de estas personas pueden representarse como en el diagrama de abajo.

Las líneas en este diagrama representan los senderos espacio-temporales que serían trazados por los cuerpos de estas personas. Llamo a cada línea simple, entre dos puntos de división, una *rama*. Y el conjunto de la estructura es el *árbol*. Cada rama corresponde a lo que se piensa como la vida de una persona. La primera persona es *Eva*. Las dos siguientes son *Segunda* y *Tercera*. La quinta persona árbol abajo es *Quincuagésima*.

Al comienzo de sus vidas, Segunda y Tercera están completamente conectadas con Eva, como ella era justo antes de dividirse, desde el punto de vista psicológico. Como he sostenido, la relación de Eva con cada una de esas dos personas es casi tan buena como



la supervivencia ordinaria. Las mismas afirmaciones se aplican a todas las demás divisiones en la historia de esta comunidad.

¿Qué deberíamos decir de la relación de Eva con personas que están más alejadas árbol abajo, como Quincuagésima?

Eva es psicológicamente continua con Quincuagésima. Entre las dos habrá una cadena continua parcialmente superpuesta de cone-

xiones psicológicas directas. De modo que Eva tiene algunas cuasi-intenciones que son realizadas por Tercera, que a su vez tiene algunas cuasi-intenciones que son realizadas por Sexta, y así sucesivamente hasta descender a Quincuagésima. Y Quincuagésima puede cuasi-recordar la mayor parte de la vida de su predecesora inmediata, que puede cuasi-recordar la mayor parte de la vida de su predecesora inmediata, y así hasta volver a Eva.

Aunque Quincuagésima es psicológicamente continua con Eva, puede que entre las dos no se dé ninguna conexividad psicológica distintiva. La conexividad requiere conexiones directas. Si estas personas son en otros aspectos como nosotros, Eva no puede estar conectada fuertemente con cada persona en un árbol que es indefinidamente largo. Con el paso del tiempo, los cuasi-recuerdos se debilitarán, y luego se desvanecerán. Las cuasi-ambiciones, una vez cumplidas, serán sustituidas por otras. Las cuasi-características cambiarán gradualmente. A causa de tales hechos, si una persona está más alejada árbol abajo, habrá menos conexiones directas entre ella y Eva. Si la persona es suficientemente remota, puede que entre las dos *no* haya conexiones psicológicas directas y distintivas. Asumamos que esto es lo que ocurre con Eva y Quincuagésima.

Escribo *distintivas* porque habría algunas clases de conexión directa. Quincuagésima heredaría de Eva muchos recuerdos de hechos, como el de que ella y las demás se reproducen por división. Y heredaría muchas destrezas generales, como hablar y nadar. Pero no heredaría ninguno de los rasgos psicológicos que distinguen a Eva de la mayoría de las demás personas de esta comunidad.

Entre Eva y Quincuagésima hay continuidad psicológica, pero no conexividad psicológica distintiva. Este caso ilustra una pregunta que mencioné antes: ¿cuál es la importancia relativa de estas dos relaciones?

Creo que ambas importan. Otros pueden pensar que una importa más que la otra, pero no conozco ningún argumento a favor de tal creencia. Asumiré que ninguna relación importa más que la otra. (Lo que no significa asumir que su importancia sea exactamente igual. Tal pregunta podría no tener una respuesta exacta.)

Puesto que será importante después, consideraré una opinión diferente. Según ella, la conexividad no importa, sólo importa la

continuidad. Si va a haber más adelante una persona que será psicológicamente continua conmigo tal y como soy ahora, no importaría en absoluto que entre esa persona y yo ahora no hubiese conexiones psicológicas directas.

Como he dicho, ciertas reducciones en conexividad podrían ser bienvenidas como mejoras. Pero no podemos afirmar justificadamente que no importaría si no hubiera ninguna conexividad psicológica. Consideremos en primer lugar la importancia de la memoria. Si nuestra vida ha valido la pena vivirla, la mayoría de nosotros valoraría enormemente nuestra capacidad de recordar muchas de nuestras experiencias pasadas. La pérdida de todos estos recuerdos no tiene por qué destruir la continuidad de la memoria, que sólo requiere cadenas de recuerdos que se solapen. Supongamos que yo sé que, dentro de dos días, mis únicos recuerdos-de-experiencia van a ser de experiencias que voy a tener mañana. Según la idea recién formulada, puesto que habrá continuidad de memoria, esto es todo lo que importa. No debería importarme que pronto vaya a perder todos mis recuerdos de mi vida pasada. La mayoría de nosotros estaría totalmente en desacuerdo. Perder todos esos recuerdos sería algo que lamentaríamos profundamente.

Consideremos a continuación la continuidad de nuestros deseos e intenciones. Supongamos que ahora quiero a determinadas personas. Podría dejar de quererlas sin ningún tipo de ruptura en la continuidad psicológica. Pero yo lamentaría enormemente esos cambios. Supongamos que también deseo intensamente alcanzar ciertos fines. Supuesto el hecho de que tengo estos intensos deseos, lamentaría su sustitución por otros. Tengo que preocuparme más *ahora* de la consecución de lo que *ahora* me preocupa. Como me preocupa más la satisfacción de los deseos presentes, lamentaría perderlos. Más en general, quiero que mi vida tenga ciertas clases de unidad global. No quiero que sea muy episódica, con fluctuaciones continuas en mis deseos e intereses. Esas fluctuaciones son compatibles con una continuidad psicológica completa, pero reducirían la conexividad psicológica. Este es otro tipo de cambio que la mayoría de nosotros lamentaría.

Consideremos finalmente la continuidad de carácter. Se dará semejante continuidad si nuestro carácter cambia de un modo natural. Pero la mayoría de nosotros valora ciertos aspectos de nuestro carácter. Querremos que estos *no* cambien. Aquí, de nuevo, queremos conexividad, no mera continuidad.

He descrito tres razones por las que la mayor parte de nosotros rechazaría la idea de que la conexividad psicológica no importa. Como son buenas razones, estas observaciones parecen suficientes para refutar esta opinión. Podemos conceder que la conexividad no es todo lo que importa. La continuidad psicológica también importa. Pero deberíamos rechazar la opinión de que sólo importa la continuidad.

101. YOES SUCESIVOS

Imaginé a personas iguales que nosotros, salvo que se reproducen por división natural. Ahora sugeriré cómo podrían describir sus interrelaciones. Cada persona es un *yo*. Eva puede pensar de cualquier persona, en cualquier lugar del árbol, que es *uno de sus yoes descendientes*. Esta frase implica continuidad psicológica dirigida al futuro. A diferencia de Eva, Tercera tiene yoes descendientes sólo en la mitad derecha del árbol. Para implicar continuidad dirigida al pasado, estas personas pueden usar la frase *un yo ancestral*. Los yoes ancestrales de Quincuagésima son todos los de las personas en la línea simple que la conecta con Eva.

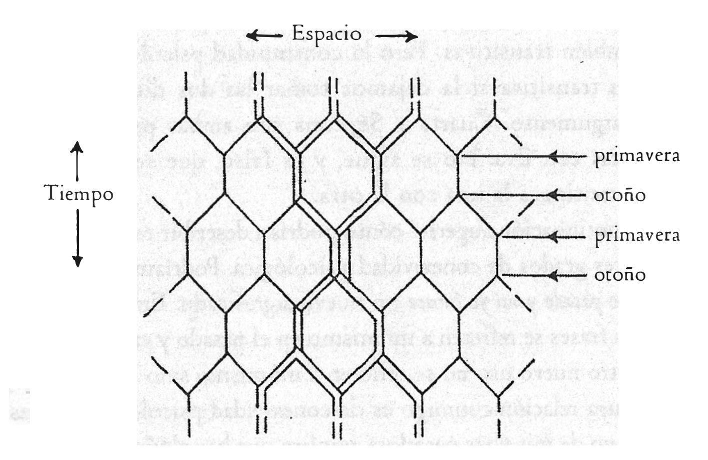
Como la continuidad psicológica es una relación transitiva, en cada dirección del tiempo, *ser un yo ancestral de* y *ser un yo descendiente de* son también transitivas. Pero la continuidad psicológica no es una relación transitiva si la dejamos tomar las dos direcciones en un único argumento. Cuarta y Séptima son ambas psicológicamente continuas con Eva. No se sigue, y es falso, que sean psicológicamente continuas la una con la otra.

A continuación sugeriré cómo podrían describir estas personas los diferentes grados de conexividad psicológica. Podríamos dar a las frases *mi yo pasado* y *mi yo futuro* un nuevo significado. En su uso corriente, estas frases se refieren a mí mismo en el pasado y en el futuro. Pero en nuestro nuevo uso no se refieren a mí mismo sino a esas otras personas cuya relación conmigo es de conexividad psicológica. Por eso la frase «uno de mis yoes pasados» implica que hay algún grado de cone-

xividad. Para implicar los diferentes grados tenemos la serie siguiente: «mi yo pasado más cercano», «uno de mis yoes pasados más cercanos», «uno de mis yoes pasados más distantes», «apenas uno de mis yoes pasados (sólo puedo cuasi-recordar unas pocas experiencias suyas)», y, finalmente, «no uno de *mis* yoes pasados, simplemente un yo ancestral». Esta es la serie de frases dirigidas al pasado que podría usar Quincuagésima. Eva podría usar una serie similar dirigida al futuro.

Este modo de hablar le vendría bien, está claro, a mi gente imaginaria. Les permitiría describir con más precisión las interrelaciones que se dan entre ellas. Este modo de hablar también proporciona una descripción nueva y plausible del caso imaginario en que me divido. Aunque yo no sobrevivo a mi división, las dos personas resultantes son dos de mis yoes futuros. Y están tan próximos a mí como yo lo estoy a mí mismo mañana. De forma parecida, cada uno de ellos puede referirse a mí como a un yo pasado igualmente próximo. (Pueden compartir un yo pasado sin ser el mismo yo que el otro.)

Consideremos a continuación otra clase de personas imaginarias. Se reproducen por fusión además de por división. Y lo hacen a menudo. Se fusionan en otoño y se dividen en primavera. Sus relaciones son como se representa abajo



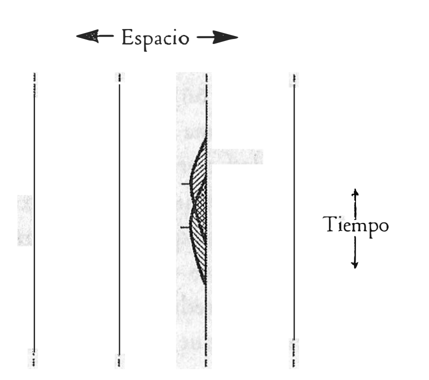
A es la persona cuya vida está representada por la rama de tres líneas. El árbol de dos líneas representa las vidas que son psicológicamente continuas con la de A. Cada persona tiene su propio árbol de dos líneas, que se superpone pero que es diferente de los árboles de las demás.

Para estas personas imaginarias, las frases «un yo ancestral» y «un yo descendiente» incluirían demasiado como para ser de mucho uso. Podría haber pares de fechas tales que todo el que vivió antes de la primera fecha fue un yo ancestral de todo el que vivirá después de la segunda fecha. Y, como la vida de cada persona dura sólo medio año, la palabra «yo» cubriría demasiado poco como para hacer todo el trabajo que hace para nosotros. Mucho de este trabajo tendría que ser hecho, para estas personas, por el habla acerca de los yoes pasados y futuros.

Hay un fallo en esta manera de hablar que se propone. La frase «un yo pasado de» implica conexividad psicológica. Y las variantes de esta frase pueden usarse para dar a entender los diversos grados de conexividad psicológica. Pero lo que distingue a los yoes sucesivos no es un grado reducido de conexividad. Los yoes se distinguen por las fusiones y divisiones. Por eso no podemos usar estas frases para dar a entender conexividad reducida dentro de una vida singular.

Este fallo no afectaría a las personas imaginarias que describí hace un momento. Ellas se dividen y se unen tan frecuentemente, y sus vidas son en consecuencia tan cortas, que dentro de una vida singular la conexividad psicológica se daría siempre en un grado muy alto.

Consideremos, finalmente, otra clase de personas imaginarias. Una vez más, esta gente difiere de nosotros sólo en su método de reproducción. Ellos *no* se reproducen. En su mundo no hay ni reproducción sexual ni división ni fusión. Hay cierto número de cuerpos que duran siempre, cambiando gradualmente de apariencia. Y las conexiones psicológicas directas y distintivas se dan, como antes, sólo a lo largo de períodos de tiempo limitados, como por ejemplo quinientos años. Esto se muestra en el diagrama siguiente



Los dos sombreados representan los grados de conexividad psicológica con sus dos puntos centrales.

Estas personas no podrían utilizar el modo de pensar que he propuesto. Como no hay ramificación de la continuidad psicológica, tendrían que considerarse a sí mismos inmortales. En un sentido, esto es lo que son. Pero deben trazar otra distinción.

Estas personas tendrían una razón para pensarse a sí mismos inmortales. Las partes de cada «línea» son todas psicológicamente continuas. Pero sólo entre las partes que están mutuamente cercanas hay conexiones psicológicas directas y distintivas. Esto da a estas personas una razón para no pensar que cada «línea» corresponde a una vida singular e indiferenciada. Si lo pensaran, no tendrían modo de dar a entender estas conexiones psicológicas directas. Cuando una persona semejante dice, por ejemplo, «Pasé una temporada explorando el Himalaya», sus oyentes no tendrían derecho a asumir que el hablante tiene recuerdos ningunos de esta temporada, o que su carácter de entonces y de ahora son de algún modo parecidos, o que ahora está realizando alguno de los planes o de las intenciones que entonces tenía. Puesto que la palabra «yo» no conllevaría ninguna de estas implicaciones, no tendría para estas personas inmortales la utilidad que tiene para nosotros.

Para dar a estas personas una mejor manera de hablar, voy a revisar mi propuesta anterior. La distinción entre yoes sucesivos puede hacerse por referencia, no a la ramificación de la continuidad psicológica, sino a los grados de conexividad psicológica. Como esta

conexividad es una cuestión de grado, el trazado de estas distinciones puede dejarse a la elección del hablante, y permitir que varíe de contexto a contexto.

Como estas distinciones se establecen ahora dentro de una vida individual, hemos vuelto mucho más cerca del uso corriente de las frases «mi yo pasado» y «mi yo futuro». Según la manera de hablar que propuse, usamos «yo», y los otros pronombres, para referirnos sólo a las partes de nuestra vida con las que, cuando hablamos, tenemos las conexiones psicológicas más fuertes. Cuando las conexiones se han reducido ostensiblemente —cuando ha habido un cambio significativo de carácter, o de estilo de vida, o de creencias e ideales— podríamos decir, «No fui *yo* el que hizo eso, sino un yo anterior». Entonces podríamos describir de qué modos, y hasta qué punto, estamos relacionados con este yo pasado.

Este modo de hablar no sólo le vendría bien a esta gente inmortal imaginaria. Es a menudo útil y natural en nuestras propias vidas. Aquí tenemos unos ejemplos de dos escritores muy diferentes:

> «Mientras estamos enamorados, somos incapaces de actuar como adecuados predecesores de las personas en que, cuando ya no lo estemos, nos convertiremos dentro de poco...» [61].

> «Nuestro pavor ante un futuro en el que tendremos que renunciar a la visión de los rostros, y al sonido de las voces que amamos, a los amigos que hoy nos otorgan nuestras alegrías más vivas, este pavor, lejos de disiparse, se intensifica si a la pena producida por una privación semejante consideramos que se añadirá lo que ahora nos parece anticipadamente una pena todavía más cruel: no sentirla como pena en absoluto —quedarnos indiferentes: porque si eso ocurriera, entonces nuestro yo habría cambiado—. Sería, en un sentido bien real, la muerte de nosotros mismos, una muerte seguida, es verdad, por una resurrección, pero en un yo diferente, cuya vida y cuyo amor quedarían más allá del alcance de esos elementos del yo existente que están condenados a morir...» [62].

[61] Proust (2), p. 226. (He alterado levemente la traducción.)
[62] Proust (2), p. 349.

«No es porque estén muertas por lo que nuestro cariño hacia ciertas personas se debilita, es porque nosotros mismos estamos muriendo. Albertine no tenía razones para reprender a su amigo. El hombre que usurpaba su nombre simplemente lo había heredado... Mi nuevo yo, mientras crecía a la sombra del viejo, con frecuencia le había oído al otro hablar de Albertine; a través de ese otro yo... pensaba que la conocía, que la encontraba atractiva... Pero se trataba sólo de un cariño de segunda mano» [63].

«Nadya había escrito en su carta: "Cuando vuelvas...". Pero en eso consistía el horror: que no habría *vuelta*... Una persona nueva, en absoluto familiar, entraría llevando el nombre de su marido, y ella vería que el hombre, su amado, por el que ella se había encerrado esperándole catorce años, ya no existía...» [64].

«Innokenty sintió lástima de ella y aceptó venir... Sintió lástima, no por la esposa con la que vivía y con la que sin embargo no vivía en estos días, sino por la muchacha rubia de los tirabuzones cayéndole sobre los hombros, la muchacha que había conocido en el décimo curso» [65].

Como sugieren estos fragmentos, el objeto de algunas de nuestras emociones puede que no sea otra persona intemporalmente considerada, sino otra persona durante un período en su vida. Aquí tenemos lo que me parece que es un ejemplo común. Puede estar claro para una pareja que se aman el uno al otro. Pero si preguntan si todavía *están enamorados* el uno del otro, tal vez encuentren esta pregunta desconcertante. Tal vez les parezca aún que están enamorados, aunque su conducta en relación con la otra persona, y sus sentimientos en presencia del otro puede que no lo confirmen. Si distinguieran entre yoes sucesivos, se podría resolver su perplejidad. Podrían ver que se aman el uno al otro, y que están enamorados de un yo anterior del otro.

El discurso sobre yoes sucesivos puede malentenderse fácilmente, o tomarse en un sentido demasiado literal. Debería compararse

[63] Proust (1), p. 249.
[64] Solzhenitsyn, p. 232.
[65] Solzhenitsyn, p. 393.

con el modo en que subdividimos la historia de una nación. Lo llamamos la historia de naciones sucesivas, como la Inglaterra Anglosajona, Medieval y Tudor [66].

Hay otro defecto en esta forma de hablar. Sólo se ajusta a casos en que hay una discontinuidad clara que marca el límite entre dos yoes. Pero puede haber grados reducidos de conexividad psicológica sin tales discontinuidades. Aunque es menos rígido que el lenguaje de la identidad, el discurso sobre yoes sucesivos no puede usarse para expresar esas suaves reducciones en los grados de conexividad. En tales casos tenemos que hablar directamente de los grados de conexividad.

Volveré ahora de los casos imaginarios a las vida real. Afirmaré que, si cambiamos nuestra idea de la naturaleza de la identidad personal, esto puede alterar nuestras creencias tanto sobre lo que es racional como sobre lo que es moralmente correcto o incorrecto.

[66] Penelhum manifiesta algunas dudas sobre este modo de hablar. Parfit (2) trata brevemente de hacer frente a estas dudas.

14

IDENTIDAD Y RACIONALIDAD

102. La Tesis Radical

Reconsideremos la teoría del Propio Interés. Afirma que, para cada persona, hay un fin último supremamente racional: que las cosas 537 marchen lo mejor posible para sí misma. Un agente racional debería tener, y además ser en último término gobernado por, una predisposición temporalmente neutral en su propio favor. Es irracional para alguien hacer lo que cree que será peor para él.

Algunos autores afirman que, si es verdadera la Concepción Reduccionista, no tenemos *ninguna* razón para estar interesados en nuestro propio futuro. A esta la llamo la *Tesis Radical*. Butler escribió que, según una versión reduccionista del punto de vista de Locke, sería «una falacia... imaginar... que nuestro yo presente estará interesado en lo que nos vaya a suceder mañana» [68]. Tomada literalmente, esta es una predicción. Pero Butler probablemente quería decir que, si es correcta la Concepción Reduccionista, no tendríamos ninguna razón para semejante interés. Sidgwick hizo declaraciones similares en relación con el punto de vista de Hume.

[68] En Perry (1), p. 102.

Según este punto de vista, «el "yo" permanente e idéntico no es un hecho sino una ficción»; el «Ego es simplemente una... serie de estados conscientes*». Sidgwick preguntó

> «¿Por qué... una parte de la serie de estados conscientes debería estar más interesada en otra parte de la misma serie en vez de estarlo más por cualquier otra serie?» [69].

Wiggins sugiere que esta pregunta no tiene respuesta [70]. Y Madell escribe

> «Es obvio que tengo todas las razones del mundo para estar preocupado por si la persona que va a tener dolor soy yo, pero no es en absoluto obvio que tenga *alguna* razón para estar interesado en el hecho de que la persona que va a tener dolor tendrá un determinado conjunto de impresiones de memoria...

añadiendo

> ... no clarifica la cuestión...que se me diga que en esta clase de contexto eso es todo lo que ser yo implica» [71].

Otros autores no tienen dudas. Según la Concepción Reduccionista, la identidad personal no consiste más que en la continuidad física y psicológica. Swinburne afirma que, si no hay nada más en la identidad personal que estas continuidades, debemos quedarnos indiferentes ante la cuestión de vivir o morir. Para decirlo con sus palabras, «en sí misma, con seguridad esa continuidad carece de valor» [72].

Swinburne rechaza la Concepción Reduccionista. Pero al menos dos reduccionistas hacen declaraciones similares. Perry afirma que,

* «Feelings» en el original, traducido como «estados conscientes» atendiendo a la sugerencia en este sentido del propio autor. (N. del t.)

[69] Sidgwick (1), pp. 418-19.

[70] Wiggins (6).

[71] Madell, p. 110.

[72] Swinburne, p. 246.

si yo simplemente sé que *alguien* va a tener dolor, tengo una razón para evitar ese dolor, si puedo. Si yo me entero de que esa persona voy a ser *yo*, la mayoría de nosotros pensaría que tengo «una razón adicional» para evitar ese dolor. Pero Perry escribe que, según su explicación reduccionista de la identidad personal, no parece haber nada que justifique esta afirmación. Tengo una razón para evitar el dolor de un completo desconocido. Y, a no ser que vaya a interferir con la realización de mis proyectos presentes, sólo tengo esta misma razón para evitar mi propio dolor futuro. Que un dolor vaya a ser *mío* no me da, en sí mismo, ninguna razón *más* para evitar el dolor [73]. Wachsberg está de acuerdo [74].

¿Debemos aceptar esta Tesis Radical? Primero yo debería hacer algún comentario adicional sobre una distinción que se hizo arriba. Una cuestión es la de cómo, en caso de que nos hiciéramos reduccionistas, afectaría esto a nuestras actitudes y a nuestras emociones. Y otra cuestión distinta es la de si, en caso de que la Concepción Reduccionista sea verdadera, estas actitudes o emociones están justificadas. Como he dicho, cuando dejé de creer en la Concepción No-Reduccionista, me volví menos preocupado por mi propio futuro. Pero aún estoy mucho más preocupado de lo que lo estaría por el futuro de un simple desconocido. Aunque estoy menos preocupado por mi futuro, si yo supiera que más tarde tendría un gran dolor, aún estaría enormemente angustiado. Si otras personas se hicieran reduccionistas, habría como mucho un efecto parecido sobre su preocupación. Esto es lo que deberíamos esperar, según cualquier concepción. Un interés especial por el propio futuro de uno sería seleccionado por la evolución. Los animales sin tal interés sería más probable que murieran antes de transmitir sus genes. Este interés persistiría, como hecho natural, aunque decidiéramos que no está justificado. Pensando a fondo los argumentos, podríamos ser capaces de suspender brevemente este interés natural. Pero pronto reviviría.

[73] «The Importance of Being Identical» [«La importancia de ser idéntico»], en Rorty, pp. 78-85.

[74] *Personal Identity, the Nature of Persons, and Ethical Theory* [*Identidad personal, la naturaleza de las personas y la teoría ética*], Tesis Doctoral, Princeton University, 1983.

Como he dicho, si una actitud tiene una explicación evolucionista, este hecho es neutral. Ni apoya ni socava la afirmación de que esta actitud está justificada. Pero hay una excepción. Puede decirse que, como todos tenemos esta actitud, hay una razón para creer que está justificada. *Esta* afirmación es socavada por la explicación evolucionista. Como existe esta explicación, todos tendríamos esta actitud aunque no estuviera justificada; de modo que el hecho de que tenemos esta actitud no puede ser una razón para creer que está justificada. Si está justificada es una cuestión abierta, que espera ser contestada.

¿Deberíamos aceptar la Tesis Radical de que, si la Concepción Reduccionista está en lo correcto, no tenemos ninguna razón para estar especialmente interesados en nuestro propio futuro? Consideremos la razón de Swinburne para afirmar esto. Swinburne afirma que, en sí mismas, las continuidades física y psicológica no tienen valor.

Una vez escribí, refiriéndome a esta y a parecidas afirmaciones:

> «Estas afirmaciones son demasiado fuertes. ¿Por qué no iban a tener las continuidades psicológicas significación racional? Hasta para la Concepción *No-Reduccionista*, seguramente les tenemos que reconocer significación. Si conserváramos nuestra identidad pero fuéramos despojados de todas las continuidades, no podríamos hacer nada en absoluto. Sin las conexiones de memoria e intención, ni podríamos actuar ni trazar planes, ni siquiera pensar» [75]

Como escribe Wachsberg [76], esta es una mala respuesta. Swinburne cree que la identidad personal es un hecho profundo adicional, distinto de la continuidad física y psicológica, y cree que este hecho adicional es lo que nos da razones para tener un interés especial por nuestro propio futuro. Yo afirmé que, sin la continuidad psicológica, no podríamos ni pensar ni actuar. Esta no es una objeción a la concepción de Swinburne. Swinburne podría conceder

[75] Parfit (6), p. 229.
[76] *Op. cit.*

que, cuando la añadimos al hecho adicional de la identidad personal, la continuidad psicológica es de gran importancia. Esto no muestra que, en la ausencia de este hecho adicional, la continuidad psicológica nos dé razones para un interés especial.

También escribí:

> «Las continuidades pueden parecer triviales cuando las comparamos con el "hecho adicional", y, con todo, ser inmensamente importantes cuando las comparamos con cualquier otro hecho. De modo que si no hay hecho adicional —si es una ilusión— las continuidades pueden tener importancia suprema. Mientras no somos reduccionistas, el hecho adicional es como el sol, resplandeciente en nuestro cielo mental. Las continuidades, en comparación, son simplemente como una luna diurna. Pero cuando nos hacemos reduccionistas, el sol se pone, y la luna puede entonces hacerse más brillante que ninguna otra cosa. Puede dominar el cielo» [77].

Swinburne podría rechazar estas declaraciones. La noche no es el día. Según la concepción de Swinburne, sólo el hecho adicional nos da razones para un interés especial. Si esta Tesis Radical está justificada, y no existe semejante hecho, no tenemos esas razones.

Puede ser útil volver al caso imaginario en que me divido. Según la Concepción No-Reduccionista, hay tres posibilidades. Yo podría ser una de las personas resultantes, la otra o ninguna. Como escribe Chisholm:

> «Cuando considero estas preguntas, veo de modo claro y distinto que las cosas siguientes son verdaderas... Las preguntas "¿Seré Izquierdo?" y "¿Seré Derecho?" tienen respuestas completamente definidas. Las respuestas serán simplemente "Sí" o "No"... En lo que quiero insistir... es en que este será el caso aunque todos nuestros criterios normales de identidad personal se vengan abajo» [78].

[77] Parfit (6), p. 230.
[78] Chisholm, pp. 188-9.

Supongamos que la opinión de Chisholm fuese verdadera. Un no-reduccionista podría decir entonces: «Si yo voy a ser Derecho, ahora tengo una razón para estar especialmente interesado en el futuro de Derecho, pero no tengo ninguna razón para estar especialmente interesado en el futuro de Izquierdo. Observaciones similares se aplican si voy a ser Izquierdo. Si no voy a ser *ninguna* de esas dos personas, no tengo razones para estar especialmente interesado en el futuro de ninguna de ellas».

Estas declaraciones asumen que, en la ausencia de identidad personal, la continuidad psicológica no proporciona ninguna razón para un interés especial. Podríamos negar esta asunción. Supongamos que voy a ser Derecho. Izquierdo no será un simple desconocido. Mi relación con Izquierdo es y será muy íntima. Podríamos decir que, comparado con el futuro de un simple desconocido, tengo razones para estar más interesado por el futuro de Izquierdo.

Un no-reduccionista podría contestar: «Después de la división, cuando yo soy Derecho, Izquierdo será alguien distinto que, al menos para empezar, es exactamente como yo. Como tú mismo dijiste, yo puedo tener razones para lamentar la existencia de Izquierdo. Aunque he sobrevivido como Derecho, la mujer que amo no lo sabría. Podría creerse la falsa afirmación de Izquierdo de que él es yo. Crea ella lo que crea, la existencia de Izquierdo interferirá con su amor por mí. Puesto que esto es así, yo podría esperar racionalmente que Izquierdo muera pronto. Como yo podría tener esta esperanza racionalmente después de la división, yo podría racionalmente tener esta esperanza ahora. Esto implica que no puedo tener una razón para estar especialmente interesado en el futuro de Izquierdo. Si tuviera esa razón, debería estar angustiado por el pensamiento de la muerte prematura de Izquierdo. Pero hemos visto que yo podría dar la bienvenida racionalmente a este suceso. Aunque Izquierdo vaya a ser psicológicamente continuo conmigo tal y como soy ahora, esta continuidad no me da aquí una razón para un interés especial».

Según la Concepción No-Reduccionista, si voy a ser *una* de las personas resultantes, puede negarse verosímilmente que tenga razones para estar especialmente interesado en *la otra*. ¿Y qué ocurre con

la posibilidad que queda, la de que no vaya a ser ninguna de esas personas? Si esto es lo que ocurrirá, es más verosímil afirmar que la continuidad psicológica me da razones para un interés especial. Debo preocuparme por las personas resultantes más de lo que me preocupo por simples desconocidos. Tengo razones para desear que al menos una de esas dos personas viva una vida completa. Esto sería mejor para la mujer que amo que el que las dos personas resultantes murieran pronto. Y esta persona resultante podría acabar mi libro inacabado, y de otras maneras realizar algunos de mis deseos.

Tenemos que admitir que esta clase de interés no es como el interés especial que tenemos en nuestro propio futuro. Es un interés por alguien distinto que puede, de varios modos, actuar de mi parte. Si un egoísta tuviera un interés así, podría considerar a esta otra persona como un mero instrumento. Supongamos que me entero de que esta persona resultante va a tener que soportar un gran dolor. ¿Debería reaccionar a esta noticia como si me hubiera enterado de que *yo* tendré que soportar un dolor semejante? Los no-reduccionistas podrían verosímilmente responder No. Pueden afirmar que, si este dolor futuro no fuese revelado a la mujer que amo, ni fuera a interferir con la finalización de mi libro, no tendría yo ninguna razón especial para preocuparme de él. Este dolor no lo sentiré yo. Si soy egoísta, mi interés por la persona que tiene dolor es sólo un interés en que, de varios modos, esta persona realice mis deseos. Si su dolor no va a interferir con su realización de mis deseos, ¿por qué debería darme razones para preocuparme?

Concluyo que, si la Concepción No-Reduccionista fuese la correcta, los no-reduccionistas podrían verosímilmente aceptar lo que llamo la Tesis Radical. Según ella, sólo el hecho adicional profundo me da una razón para estar especialmente interesado en mi futuro. En ausencia de este hecho, la continuidad psicológica no me da tal razón.

Supongamos a renglón seguido que un no-reduccionista dejara de creer en su concepción. ¿Podría aceptar todavía la Tesis Radical? Si la identidad personal no implica el hecho adicional profundo, sino que consiste nada más que en la continuidad física y psicoló-

gica, mi relación con cada una de las personas resultantes es tan buena como la supervivencia corriente. Mi relación con cada una de las personas resultantes es la relación R con su causa normal, suficiente continuidad física. Si nos hemos hecho reduccionistas, debemos aceptar mi afirmación de que la relación R es tan buena como la supervivencia corriente. Pero esta afirmación no implica que, cuando R se da, nos aporte una razón para estar especialmente interesados en nuestro propio futuro. Podríamos aceptar la Tesis Radical de que, si la supervivencia corriente no implica el hecho adicional profundo, no nos aporta esa razón.

Supongamos que aceptamos otra de mis afirmaciones: que no importaría que la continuidad psicológica tuviera una causa anormal. Entonces concederíamos que, en el teletransporte, mi relación con mi Réplica es tan buena como la supervivencia corriente. Pero desde el punto de vista de nuestra anterior Concepción No-Reduccionista, esta afirmación estaría mejor si la pusiéramos de un modo diferente. Como escribí, la supervivencia corriente *no es mejor que*, o es *tan mala como*, mi relación con mi Réplica. Y yo podría decir ahora: «Mi relación con mi Réplica no me aporta ninguna razón para interesarme especialmente. Como la supervivencia corriente no es mejor, tampoco me aporta tal razón».

Esta línea de razonamiento es justificable. Cuando un no-reduccionista deja de creer que la identidad personal implica el hecho adicional profundo, puede mantener con justificación su idea de que sólo este hecho nos aportaría razones para interesarnos especialmente. Puede aceptar la Tesis Radical de que, si no existe semejante hecho, no tenemos esas razones. Y podríamos aceptar justificadamente esta afirmación aunque hayamos sido siempre reduccionistas.

¿Podría cambiar justificadamente un no-reduccionista su parecer? ¿Podría afirmar que la relación R nos aporta una razón para interesarnos especialmente? Llamo a esta la *Tesis Moderada*.

Creo que, al igual que la radical, esta tesis es justificable. No conozco un argumento que demuestre que, de estas dos tesis, es la moderada la que debemos aceptar. Podría decirse

Los radicales se equivocan al dar por sentado que sólo el hecho adicional profundo nos aporta una razón para interesarnos especialmente. Pensemos en nuestro interés especial en nuestros propios hijos, o en alguien a quien queremos. Dada la naturaleza de nuestra relación con nuestros hijos, o con alguien a quien queremos, podemos afirmar con justificación que tenemos razones para estar especialmente interesados en lo que vaya a ocurrirles a estas personas. Y las relaciones que justifican este interés especial no son el hecho separado profundo de la identidad personal. Si estas relaciones nos aportan una razón para interesarnos especialmente, podemos decir lo mismo de la relación R. Podemos decir que esta relación nos aporta a cada uno de nosotros una razón para estar especialmente interesados en nuestro propio futuro.

Pero un radical podría contestar:

¿Por qué debería preocuparme lo que les vaya a ocurrir más adelante a las personas que quiero? La razón no puede ser porque yo las querré todavía más adelante. Esto no es dar ninguna respuesta, porque nuestra pregunta es por qué debería preocuparme ahora, especialmente, lo que vaya a preocuparme más adelante. Ni tampoco podría la razón ser, «Porque a los que quiero les preocupa ahora lo que vaya a ocurrirles más adelante». Esto no es dar ninguna respuesta, porque nuestro problema es también saber por qué *ellos* deberían preocuparse de lo que les vaya a ocurrir más adelante. Todavía no tenemos respuesta a la pregunta de por qué, en ausencia del hecho adicional profundo, deberíamos estar especialmente interesados en nuestro propio futuro o en el futuro de alguien [79].

Esta objeción tiene alguna fuerza. Y puede ser erróneo comparar nuestro interés en nuestro propio futuro con nuestro interés por aquellos a los que queremos. Supongamos que me entero de que alguien que quiero sufrirá pronto un gran dolor. Me angustiará tremendamente esta noticia. Podría angustiarme *más* que si me entera-

[79] Esta respuesta me la sugirió J. Broome.

ra de que *yo* pronto voy a sufrir ese dolor. Pero esta preocupación tiene una cualidad diferente. Yo no *anticipo* el dolor que sentirá la persona que quiero. Podría decirse que sólo según la Concepción No-Reduccionista podemos anticipar justificadamente dolores futuros. La anticipación podría justificarse sólo por el no existente hecho adicional profundo. Tal vez, si somos reduccionistas, deberíamos dejar de anticipar nuestros propios dolores futuros [80].

Si esta última afirmación es verdadera, aporta una razón suplementaria para pensar que, según la Concepción Reduccionista, no tenemos ninguna razón para estar especialmente preocupados por nuestro propio futuro. Pero esta afirmación no nos fuerza a aceptar esta conclusión. Parece defendible tanto afirmar cuanto negar que la relación R nos aporta una razón para preocuparnos especialmente. Aunque no estemos forzados a aceptar la Tesis Radical, podemos ser incapaces de demostrar que debería ser rechazada. Hay una gran diferencia entre las Tesis Radical y Moderada. Pero todavía no he encontrado un argumento que refute una de las dos.

¿Qué consecuencias tienen estas conclusiones para la teoría del Propio Interés sobre la racionalidad? La Tesis Radical es la negación
546 radical de esta teoría. Este puede ser el argumento contra PI que, como dije en la Sección 54, Sidgwick medio sugirió.

Como la Tesis Radical es defendible, este argumento consigue algo. Pero la Tesis Radical puede también negarse con justificación. Por tanto, el argumento no refuta la teoría del Propio Interés.

103. Un argumento mejor contra PI

Lo podemos hacer mejor. He estado afirmando

(A) Como la identidad personal no implica el hecho adicional profundo, es menos profunda, o implica menos.

También he defendido

[80] Esto lo sugiere Wachsberg.

(B) Lo que fundamentalmente importa son la conexividad y la continuidad psicológicas.

La tesis radical, que apela a (A), puede ser negada. Pero (B) aporta la premisa para un nuevo desafío a la teoría del Propio Interés.

Es central a esta teoría

El Requisito de la Igual Preocupación: una persona racional debería estar *igualmente* preocupada por *todas* las partes de su futuro.

Como escribe Sidgwick, «mis estados conscientes de aquí a un año deberían ser igual de importantes para mí que mis estados conscientes del minuto próximo, con tal de que pudiese hacer un pronóstico de los mismos igualmente seguro. Verdaderamente, esta preocupación igual e imparcial por todas las partes de la vida consciente de uno es tal vez el elemento más prominente en la noción común de lo *racional*...» [81]. Cada uno de nosotros puede racionalmente dar menos importancia a lo que puede ocurrir más adelante en el futuro, si este alejamiento hace al suceso menos probable. 547
Pero, al parecer de la teoría del Propio Interés, no podemos racionalmente preocuparnos menos por nuestro futuro más distante simplemente porque sea más lejano. PI afirma, por tanto, que es irracional posponer una terrible experiencia si uno sabe que esto la va a hacer peor.

Apelando a la concepción reduccionista, podemos poner en duda esta última afirmación. Para simplificar nuestro desafío, podemos asumir que la mera proximidad temporal no puede importar. Podemos asumir que es irracional cuidarse menos de nuestro futuro más distante simplemente porque está más lejos en el futuro. Esto no demuestra que sea irracional cuidarse menos por nuestro futuro más distante. Puede haber otra razón para hacerlo así.

Como he sostenido, lo que fundamentalmente importa son la conexividad y la continuidad psicológicas. También afirmé que estas

[81] Sidgwick (1), p. 124.

importan como quiera que vengan causadas. Dado que ahora estamos considerando nuestra vida real, esta segunda afirmación es irrelevante. Nuestra afirmación podría ser que estas dos relaciones son lo que importa, contando con que tengan su causa normal.

También argüí, en la Sección 100, que *las dos* relaciones importan. No podemos afirmar justificablemente que sólo importe la continuidad. Tenemos que admitir que la conexividad importa.

Como esta relación importa, afirmo

(C) Mi preocupación por mi futuro puede corresponder al grado de conexividad entre yo ahora y yo mismo en el futuro. La conexividad es una de las dos relaciones que me aportan razones para estar especialmente interesado en mi propio futuro. Puede ser racional preocuparme menos, cuando uno de los fundamentos de la preocupación se dará en un grado menor. Como la conexividad es casi siempre más débil en períodos más largos, puedo racionalmente preocuparme menos por mi futuro más distante.

Esta afirmación defiende una nueva clase de tasa de descuento.

Se trataría de una tasa de descuento no con respecto al tiempo mismo, sino con respecto al debilitamiento de una de las dos relaciones que son lo que fundamentalmente importa. A diferencia de la tasa de descuento con respecto al tiempo, esta nueva tasa de descuento se aplicará rara vez sobre el futuro cercano. Las conexiones psicológicas entre yo ahora y yo mismo mañana no son mucho más íntimas que las conexiones entre yo ahora y yo mismo el mes próximo. Y puede que no sean mucho más íntimas que las conexiones entre yo ahora y yo mismo el año que viene. Pero sí son mucho más íntimas que las conexiones entre yo ahora y yo mismo dentro de cuarenta años.

Una vez defendí (C) apelando a

(D) Cuando una relación importante tiene lugar en un grado diferente, no es irracional creer que tiene un grado de importancia diferente.

Afirmé que, como no podemos negar (D) justificadamente, tenemos que aceptar (C) [82]. Como puntualiza Kagan, este es otro argumento malo. Una apelación a (D) no puede dar apoyo a (C). Podría afirmarse

(E) Hay como mínimo una excepción a (D). Quizás podamos imaginar casos en que un gran debilitamiento de las conexiones justificaría una menor preocupación. Pero nunca hay, en efecto, en las vidas corrientes, un gran debilitamiento semejante. En los casos reales, aunque la conexividad se vaya a dar en grados reducidos, debemos considerarla como teniendo la *misma* importancia. Es irracional preocuparse menos por nuestro futuro más distante, simplemente porque vaya a haber tal reducción en el grado de conexividad.

Si (E) se puede defender, (D) no es verdadero. Por tanto, no podemos rechazar (E) con una apelación a (D). Tal apelación daría por supuesto lo que hay que demostrar. No podemos argumentar a favor de una conclusión apelando a una afirmación que asume esta conclusión. (Puede ser útil dar un ejemplo. Supongamos que hay

muchas canicas en una caja. Yo he cogido todas a excepción de la última, que es negra. No puedo decir, «Esta canica tiene que ser blanca, puesto que todas son blancas». Si hay una excepción a una afirmación general, no podemos negar esta excepción simplemente apelando a esta afirmación.)

Aunque no puedo apelar a (D), puedo estar en lo cierto rechazando (E). Aunque haya algunas excepciones, hay muchísimas relaciones de las que podemos creer que son menos importantes cuando se verifican en grados reducidos. Algunos ejemplos los tenemos en la amistad, la complicidad, la relevancia, el endeudamiento, ser un pariente cercano de y ser responsable por. Puede estar justificado pensar lo mismo sobre la conexividad psicológica.

Afirmo no sólo que esta creencia se puede justificar, sino también que *no puede negarse justificadamente*. Supongamos que tendré un

[82] Parfit (6), p. 232. (Mi versión de (D) era, de un modo trivial, inferior.)

día de dolor mañana, y también dentro de cuarenta años. Estoy fuertemente conectado conmigo mismo mañana, desde el punto de vista psicológico. Habrá mucha menos conexividad entre yo ahora y yo mismo dentro de cuarenta años. Como la conexividad es una de mis dos razones para preocuparme de mi futuro, no puede ser irracional que yo me preocupe menos cuando vaya a haber mucha menos conexividad.

Como (E) no se puede justificar, deberíamos aceptar (C). Al aceptar (C), estamos rechazando el requisito de igual preocupación, que es central en la teoría del Propio Interés. Por eso deberíamos rechazar esta teoría.

104. El contraargumento del teórico PI

Un teórico del Propio Interés podría apelar ahora a

> *La Perogrullada*: Todas las partes del futuro de una persona son *por igual* partes de su futuro.

Este teórico podría afirmar que, por consiguiente, es irracional no preocuparse por igual de todas las partes de nuestro futuro.

Este argumento asume que la identidad personal es lo que importa. Cuando consideramos el caso imaginario en que me divido, nos damos cuenta de que la identidad personal no es lo que importa. Aunque yo no seré ninguna de las dos personas resultantes, mi relación con cada una de ellas contiene lo que importa.

Como el contraargumento del teórico PI asume de manera falsa que la identidad es lo que importa, no necesitamos discutirlo. No demuestra nada un argumento con una premisa falsa. Pero vale la pena señalar que, aunque concedamos esta premisa, el argumento del teórico PI fracasa.

Asumamos, falsamente, que la identidad personal es lo que importa. ¿Una apelación a la Perogrullada proporciona un buen argumento a favor del requisito de igual preocupación? La totalidad

del futuro de una persona es por igual su futuro. ¿Demuestra esto que esta persona deba estar ahora igualmente preocupada por la totalidad de su futuro?

Este sería un buen argumento si la Concepción No-Reduccionista fuese verdadera. Según esa concepción, la Perogrullada es una profunda verdad, lo suficientemente profunda como para dar apoyo al argumento.

Consideremos

> (F) Todos los parientes de una persona son *por igual* sus parientes.

Hay un sentido en que esto es cierto. Podemos usar «pariente de» en un sentido que carece de grados. En este uso, mis hijos y mis primos lejanos son en la misma medida mis parientes. ¿Estamos ante una verdad profunda?

Tenemos que distinguirla de otra verdad, que requiere el mismo uso de estas palabras. Según este uso, *pariente de* es una relación transitiva: los parientes de mis parientes tienen que ser mis parientes. Este es un uso útil. Desde Darwin, da una nueva significación a la Gran Cadena del Ser. Como ahora sabemos, los pájaros que se ven desde mi ventana son, en un sentido literal, mis parientes. Son mis parientes en el *mismo* sentido en que mis primos son mis parientes. Tenemos un antepasado común. Los pájaros son mis *enésimos* primos *m* veces segundos. (*Pariente de* cruza los límites entre diferentes especies. Si no lo hiciera, no podría haber evolución.)

Que todos los animales superiores sean *literalmente* mis parientes es una profunda verdad. Pero ¿es profundamente verdadero que todos sean *por igual* mis parientes —que los pájaros sean tan parientes míos como mis propios hijos?—. Esta no es una profunda verdad. Es superficial, y —aunque de hecho nunca engaña— engañosa. Que sea en absoluto verdadera es el precio que tenemos que pagar por la transitividad de *pariente de*. Supongamos que decimos, «Por "pariente" queremos decir en realidad "pariente no muy lejano"—primos décimos diez veces segundos no son en realidad parientes»—. Esto nos despojaría de la profunda verdad de que los pájaros son literalmente mis parientes. Para preservar esa verdad

tenemos que conceder que —en un sentido superficial— los pájaros son *tan* parientes míos como mis propios hijos.

Como es superficial, (F) no puede dar apoyo a la clase de argumento que estamos considerando. Supongamos que, creyendo firmemente en los vínculos de parentesco, dejo todo mi dinero a mis diversos parientes. Anuncio mis intenciones mientras estoy vivo. Tengo la intención de dejar la parte más sustanciosa a mis propios hijos. ¿Podrían mis primos apelar verosímilmente a (F)? ¿Podrían argumentar que, como ellos son *igualmente* mis parientes, ellos (y los pájaros) deberían tener partes iguales? Evidentemente no. Aunque es cierto que también son mis parientes, esta verdad es demasiado trivial como para dar apoyo a su argumento.

Observaciones similares se aplican a

(G) Todos los dolores son *igualmente* dolores.

Podemos usar la palabra «dolor» de un modo que haga esto verdadero. ¿Podríamos argumentar que, puesto que (G) es verdadero, es irracional preocuparse más por los dolores más intensos? Evidentemente no. Esta es otra verdad demasiado trivial como para apoyar tal argumento. Consideremos, finalmente,

(H) Todas las partes de la historia de una nación son *igualmente* partes de la historia de esa nación.

Todas las partes de la historia de Inglaterra son igualmente partes de la historia de Inglaterra. La Inglaterra Tudor es así Inglaterra. También lo fue la Inglaterra Sajona. Y también, si decidimos llamarla «Inglaterra», la Inglaterra Romana. Pero si la llamamos «Bretaña Romana» no fue Inglaterra en absoluto. Esto muestra que (H) es trivial.

Una nación es en muchos aspectos distinta de una persona. A pesar de esas diferencias, la identidad de las personas a través del tiempo es, en sus rasgos fundamentales, como la identidad de las naciones a través del tiempo. Ambas consisten en nada más que el darse a través del tiempo de varias conexiones, algunas de las cuales son cuestión de grado. Es cierto que cuando sea viejo también seré

yo. Pero esta verdad puede ser perfectamente comparada con la verdad de que (digamos) la Austria moderna es todavía *igualmente* Austria. Un descendiente de los emperadores Habsburgo podría con razón llamar a esta verdad trivial.

En esta sección he discutido el contraargumento del teórico PI. Como este argumento asume de manera falsa que la identidad personal es lo que importa, podría haber sido descartado en seguida. Pero valió la pena demostrar que, incluso si hacemos esta falsa asunción, el argumento fracasa. El argumento apela a la afirmación de que todas las partes de nuestro futuro son *por igual* partes de nuestro futuro. Esta verdad es demasiado trivial para dar apoyo al argumento. Según un uso de «pariente», es también cierto que mis hijos y mis primos son por igual mis parientes. Si mis primos argumentan que, puesto que esto es así, ellos y mis hijos deberían heredar partes iguales, deberíamos rechazar su pretensión. Por la misma razón, aunque asumamos falsamente que la identidad personal es lo que importa, deberíamos rechazar el contraargumento del teórico PI [83].

105. La derrota de la teoría clásica del propio interés

Volvamos a mi argumento en contra de la teoría del Propio Interés. Este argumento demuestra, pienso yo, que tenemos que rechazar el requisito de igual preocupación. Según este, yo debería preocuparme ahora por igual de todas las partes de mi futuro. Es irracional preocuparme menos de mi futuro más distante —tener lo que los economistas llaman una tasa de descuento—. Quizás sea esto irracional si tengo una tasa de descuento con respecto al tiempo. Pero no es irracional si tengo una tasa de descuento con respecto a los grados de conexividad psicológica.

[83] (Nota añadida en 1985.) Como B. Garrett ha señalado, en esta Sección se incurre en un error. En el caso del parentesco, lo que importa no es la relación transitiva *pariente de*, sino la relación intransitiva *pariente cercano de*, que puede tener grados. Por consiguiente no es esta una buena analogía, si asumimos que lo que importa es la identidad personal. Para hacer frente al Contraargumento, puede que tengamos que poner en cuestión esta asunción.

El teórico del Propio Interés podría revisar su punto de vista. Según la Teoría Revisada, el interés dominante de una persona racional debería ser su propio futuro, pero ahora puede estar menos preocupada por aquellas partes de su futuro con las que esté ahora menos íntimamente conectada. Esta Teoría Revisada incorpora mi nueva tasa de descuento. Según esta teoría, no estamos racionalmente requeridos a tener esta tasa de descuento. Pero si la tenemos no somos irracionales.

Esta revisión marca una gran diferencia. Rompe la conexión entre la teoría del Propio Interés y lo que va a favor de los mejores intereses de uno. Según la Teoría no revisada, o *Clásica*, es irracional para cualquiera hacer lo que cree que va a ser peor para él. Según la Teoría Revisada del Propio Interés, esta afirmación tiene que abandonarse. Si no es irracional interesarse menos por algunas partes de nuestro futuro, puede que no sea irracional hacer lo que uno cree que va a ser peor para sí mismo. Puede que no sea irracional actuar a sabiendas contra nuestro propio interés.

Como muestra esta última afirmación, la Teoría Revisada no es una versión de la teoría del Propio Interés. Es una versión de la teoría Crítica del fin Presente. Pero es lo de menos cómo clasificamos esta teoría.

Lo que es importante es que tenemos que abandonar la tesis central de la Teoría Clásica. Consideremos la imprudencia grave y deliberada. Para disfrutar de pequeños placeres en mi juventud, me condeno a sufrir enormemente cuando sea mayor. Podría, por ejemplo, empezar a fumar cuando soy un muchacho. Sé que probablemente me voy a imponer a mí mismo una muerte prematura y dolorosa. Sé que estoy haciendo lo que probablemente va a ser mucho peor para mí. Como tenemos que rechazar la Teoría Clásica, no podemos afirmar que todos estos actos son irracionales.

Según la Teoría Revisada, tales actos *podrían* ser irracionales. Según esta teoría, no es irracional tener una tasa de descuento con respecto a los grados de conexividad psicológica. Cuando me impongo a mí mismo un gran sufrimiento en mi vejez, por disfrutar ahora de pequeños placeres, mi acto es irracional sólo si mi tasa de descuento es *demasiado pronunciada*.

Una debilidad de la Teoría Revisada es su necesidad de explicar qué hace a una tasa de descuento demasiado pronunciada. Pero el punto importante es que, aunque esta tasa *no* sea demasiado pronunciada, todos estos actos tienen que ser criticados. Una gran imprudencia es siempre lamentable, y con frecuencia (como en el caso de fumar) trágica. Según la Teoría Revisada, no podemos afirmar que todos estos actos son irracionales. Como deberíamos criticarlos, tenemos que apelar a otra teoría.

106. La inmoralidad de la imprudencia

¿Cómo deberíamos criticar la imprudencia grave? Podríamos decir que, simplemente, podemos llamar imprudentes a actos así. Podría decirse que esto es una crítica, aunque ya no pensemos que la imprudencia sea irracional.

Muchos rechazarían esta afirmación. Consideremos la afirmación de que alguien carece de *castidad*. Mucha gente piensa ahora que no hay nada moralmente incorrecto en la falta de castidad. Y, para estas personas, la acusación de «impúdico» deja de ser una crítica. Una afirmación parecida se aplica a la acusación de «imprudente». Igual que «impúdico» expresa una objeción moral, «imprudente» expresa una objeción relacionada con la racionalidad. Esto se muestra por las palabras más comunes con las que la gente es criticada por actuar imprudentemente. Cuando alguien hace lo que sabe que va a ser peor para él, muchos lo llamarían «estúpido», «idiota», «bobo», o «tonto». Esto muestra que esta objeción tiene que ver con la irracionalidad. Si pensamos que un acto imprudente no es irracional, la acusación de «imprudente» podría, para algunos de nosotros, dejar de ser una crítica. Podría convertirse, como «impúdico», simplemente en una descripción.

Pero la imprudencia grave debe ser criticada. ¿Qué clase de crítica podemos hacer? Podría sugerirse que podemos apelar a la teoría Crítica del fin Presente. Como escribí en la Sección 52, podemos afirmar que, en nuestra preocupación por nuestro propio interés, es irracional no ser temporalmente neutral. Según esta ver-

sión de CP, la imprudencia grave es irracional. Más exactamente, es irracional a no ser que aporte a los demás grandes beneficios, o realice algún deseo que no es irracional.

Esta sugerencia falla. En mi último argumento contra la teoría del Propio Interés, *asumí* que es irracional no ser temporalmente neutral. El argumento defendía una tasa de descuento, pero no con respecto al tiempo sino con respecto a los grados de conexividad psicológica. Como esta conexividad es una de mis dos razones para preocuparme por mi futuro, no puede ser irracional para mí preocuparme menos cuando vaya a haber menos conexividad. La teoría Crítica del fin Presente no puede, justificablemente, negar esta afirmación.

La objeción a la imprudencia grave tiene que venir de otra dirección. Sugiero que, como debemos rechazar la Teoría Clásica del Propio Interés, deberíamos expandir el área cubierta por la moralidad. Nuestra teoría moral debería anexionarse el territorio que la Teoría Revisada del Propio Interés ha abandonado.

Como afirmaban los críticos de Mill, los actos *puramente* «egoístas» son raros. Si yo soy enormemente imprudente, es probable que esto sea malo para otras personas. Pero, si los efectos principales de mi acto van a recaer en mí mismo, la mayor parte de nosotros no lo juzgaría moralmente incorrecto. Las versiones más antiguas de la Moralidad del Sentido Común incluyen algunos deberes hacia uno mismo. Pero son deberes especiales, como el de desarrollar el propio talento, o el de preservar la propia pureza. Rara vez se dice que la imprudencia grave sea moralmente incorrecta. Mientras se pensaba que estos actos eran irracionales, no se tenía necesidad de pensar que fuesen inmorales. Pero, como ahora tenemos que abandonar la Teoría Clásica del Propio Interés, deberíamos ampliar nuestra teoría moral.

Hay dos modos de hacerlo. Podríamos apelar al Consecuencialismo. En particular, podríamos apelar a un principio de beneficencia imparcial o neutral respecto al agente. Supongamos que, para conseguir ahora beneficios menores, me impongo a mí mismo cargas mayores en la vejez. Aquí hago lo que, considerado imparcialmente, tiene peores efectos, o incrementa la suma de sufrimiento.

Podríamos decir que mi acto es moralmente incorrecto porque incrementa la suma de sufrimiento, aun cuando vaya a ser *yo* el que sufrirá más. Más en general, mi imprudencia es incorrecta porque produzco la peor consecuencia. No sirve de excusa el que la consecuencia vaya a ser peor sólo para mí.

Podríamos ampliar también la parte de nuestra teoría que es relativa al agente. Esta parte incluye nuestras obligaciones especiales para aquellos con los que estamos en determinadas relaciones, como nuestros padres, nuestros hijos pequeños, alumnos, pacientes, clientes o electores. Una persona se halla en otra relación especial consigo misma en el futuro, una relación de la que podemos decir que crea obligaciones especiales similares.

Si revisáramos nuestra concepción moral de alguno de estos modos, esto significaría, para muchos, un gran cambio en su concepción de la moralidad. Estas personas son de la opinión de que no puede ser una cuestión moral cómo afecta uno a su propio futuro.

Tal vez sea más fácil llegar a pensar así si subdividimos la vida de una persona en la de los yoes sucesivos. Como he dicho, esto ha parecido natural desde hace mucho tiempo, cuando se da un debilitamiento pronunciado en la conexividad psicológica. Después de un debilitamiento semejante, mi yo anterior puede parecerme ajeno ahora. Si fracaso a la hora de *identificarme* con ese yo anterior, en algunos aspectos estoy pensando en ese yo como en una persona diferente.

Podríamos hacer afirmaciones similares respecto de nuestros yoes futuros. Si ahora nos preocupamos poco de nosotros mismos en el futuro más distante, nuestros yoes futuros serán como las generaciones futuras. Podemos afectarlas para peor, y, puesto que ahora no existen, no pueden defenderse a sí mismas. Como las generaciones futuras, los yoes futuros no tienen voto, de modo que sus intereses necesitan ser especialmente protegidos.

Volvamos a considerar al muchacho que empieza a fumar, sabiendo, pero sin que le importe apenas, que esto le puede hacer sufrir enormemente cincuenta años después. Este muchacho no se identifica con su yo futuro. Su actitud hacia él es en algunos aspectos como su actitud hacia otras personas. Esta analogía hace más

fácil pensar que su acto es moralmente incorrecto. Corre el riesgo de imponerse a sí mismo una muerte prematura y dolorosa. Deberíamos afirmar que es incorrecto imponerle a *cualquiera*, incluyendo ese yo futuro, el riesgo de una muerte como esa. Más en general, deberíamos afirmar que la imprudencia grave es moralmente incorrecta. No debemos hacerles a nuestros yoes futuros lo que sería incorrecto hacerles a otras personas.

15

IDENTIDAD PERSONAL Y MORALIDAD

Si nos hacemos reduccionistas, ¿deberíamos introducir otros cambios en nuestras ideas morales?

107. AUTONOMÍA Y PATERNALISMO

Somos paternalistas cuando hacemos que alguien actúe a favor de sus intereses. Proporciona cierta justificación para el paternalismo, cuando implica coacción o la violación de la autonomía de alguien, el que evitemos que esta persona actúe irracionalmente. Esto es lo que creemos que estamos haciendo si aceptamos la teoría del Propio Interés. Argumenté que tenemos que rechazar esta teoría. Pero debemos ampliar nuestra teoría moral para que incluya lo que hemos rechazado. Deberíamos afirmar que la imprudencia grave es moralmente incorrecta.

Esta afirmación refuerza los argumentos a favor de la intervención paternalista. La persona a la que coaccionamos podría decir: «Puedo estar actuando irracionalmente, pero, aun así, esto es asunto mío. Si me estoy haciendo daño sólo a mí mismo, tengo el derecho de actuar irracionalmente, y tú no tienes derecho a detenerme».

Esta respuesta tiene cierta fuerza. No pensamos tener el derecho general de evitar que la gente actúe *irracionalmente*. Pero pensamos tener el derecho general de evitar que la gente actúe *incorrectamente*. Puede que esta afirmación no se aplique a la maldad menor. Pero pensamos que no puede ser incorrecto, y que a menudo sería nuestro deber, evitar que los demás hagan lo que es gravemente incorrecto. Como debemos pensar que la imprudencia grave es gravemente incorrecta, debemos pensar que deberíamos evitar tal imprudencia, aunque ello implique coacción. La autonomía no incluye el derecho a imponerse a sí mismo, por ninguna razón válida, un gran daño. Debemos evitar que nadie haga a su yo futuro lo que sería incorrecto hacer a otras personas.

Aunque estas afirmaciones presten apoyo al paternalismo, quedan objeciones bien conocidas. Es mejor que cada uno de nosotros aprenda de sus propios errores. Y es más difícil para los demás saber que son errores.

108. Los dos extremos de la vida

Hay muchas otras maneras en que, si hemos cambiado de opinión en lo que respecta a la identidad personal, esto puede justificar un cambio en nuestras ideas morales. Un ejemplo es nuestro punto de vista sobre la moralidad del aborto. Según la Concepción No-Reduccionista, puesto que mi existencia es todo-o-nada, tiene que haber habido un momento en que yo empecé a existir. Como en mi espectro imaginario, tiene que haber una línea divisoria nítida. Es inverosímil afirmar que esta línea divisoria es el nacimiento; ni tampoco podemos trazar plausiblemente ninguna línea durante el embarazo. Puede que de este modo seamos llevados a la convicción de que yo empecé a existir en el momento de la concepción. Podemos afirmar que este es el momento en que mi vida empezó. Y, a juicio de la Concepción No-Reduccionista, es una *profunda* verdad que todas las partes de mi vida son *por igual* partes de mi vida. Yo era en la misma medida yo aun cuando mi vida hubiera acabado de empezar. Matarme en ese momento es, sencillamente, matar a una

persona inocente. Si es así como pensamos, afirmaremos plausiblemente que el aborto es moralmente incorrecto.

Según la Concepción Reduccionista, no creemos que en todo momento yo exista o no exista. Ahora podemos negar que un óvulo fertilizado sea una persona o un ser humano. Esto es como negar que una bellota sea una encina. Dadas las condiciones adecuadas, una bellota se convierte lentamente en una encina. Esta transición lleva tiempo, y es una cuestión de grado. No hay línea divisoria nítida. Deberíamos afirmar lo mismo de las personas y los seres humanos. Entonces podremos plausiblemente adoptar una perspectiva diferente sobre la moralidad del aborto. Podremos pensar que no hay nada incorrecto en un aborto temprano, pero que sería gravemente incorrecto hacer abortar a un niño próximo al término del embarazo. Ese niño, si no es deseado, debería nacer y ser dado en adopción. Los casos que se sitúan en el medio los podemos tratar como cuestiones de grado. El óvulo fertilizado no es al principio un ser humano, ni una persona, pero lentamente se convierte en eso. Del mismo modo, la destrucción de ese organismo no es al principio gravemente incorrecta, pero lentamente se convierte en eso. Después de no ser de ningún modo incorrecta, se convierte en una maldad menor, que estaría justificada en resumidas cuentas sólo si el posterior nacimiento del niño fuese mucho peor para sus padres o para otras personas. Cuando el organismo se hace del todo un ser humano, o una persona, la maldad menor se cambia en un acto que sería gravemente incorrecto.

He descrito las dos concepciones principales que son ampliamente mantenidas sobre la moralidad del aborto. La primera es apoyada por la Concepción No-Reduccionista acerca de la naturaleza de las personas, y la segunda por la Concepción Reduccionista. Aunque no es el único modo de pensar que es compatible con el Reduccionismo, creo que deberíamos adoptar esta segunda concepción.

Dentro de ella, hay espacio para el desacuerdo. La mayoría de nosotros no distingue entre personas y seres humanos. Pero algunos, siguiendo a Locke, sí que los distinguen. Éstos afirman típicamente que un ser humano llega a ser una persona sólo cuando se hace autoconsciente. Un feto se convierte en ser humano antes del

final del embarazo. Pero un niño recién nacido no es autoconsciente. Si hacemos esta distinción, podemos llegar a pensar que, mientras que es malo matar a un ser humano, es peor matar a una persona. Podemos incluso llegar a pensar que sólo es incorrecto matar a personas. No me meteré en este debate. Lo que asumen ambos bandos lo cuestionaré en la Cuarta Parte.

Consideremos a continuación el otro extremo de la vida. Para la Concepción No- Reduccionista, toda persona tiene que estar viva o muerta. Para la Concepción Reduccionista, una persona puede dejar de existir gradualmente algún tiempo antes de que su corazón deje de latir. Ocurrirá así si los rasgos distintivos de la vida mental de una persona desaparecen gradualmente, cosa que con frecuencia ocurre. Podemos afirmar de modo verosímil que, si la persona ha dejado de existir, no tenemos ninguna razón moral para ayudar a que su corazón siga latiendo, o para abstenernos de evitarlo.

Esta afirmación distingue a la persona del ser humano. Si sabemos que un ser humano está en un coma que es incurable —que este ser humano, con toda seguridad, nunca recobrará la conciencia— pensaremos que la persona ha dejado de existir. Como hay un cuerpo humano vivo, todavía existe el ser humano. Pero, en este extremo de la vida, deberíamos afirmar que sólo es incorrecto matar a las personas.

109. MERECIMIENTOS

Algunos autores afirman que, si la concepción reduccionista es verdadera, no podemos merecer ser castigados por nuestros delitos. Butler escribe que, según una versión reduccionista de la concepción de Locke, sería «caer en una falacia si acusáramos a nuestros yoes presentes de algo que hicimos...» [86]. Otro de los críticos de Locke del siglo XVIII hace una afirmación más radical. Reid contrasta la identidad personal con la identidad de cosas tales como los barcos o los árboles. La identidad de cosas como estas, escribe:

[86] En Perry (1), p. 102.

> «no es identidad perfecta; es más bien algo que, por la conveniencia del discurso, llamamos identidad. Admite un gran cambio del sujeto, siempre que sea gradual; a veces hasta un cambio total. Y los cambios que en el lenguaje común se hacen consistentes con la identidad no difieren en clase de los que se piensa que la destruyen, sino en número y en grado. La *identidad* no tiene una naturaleza fija cuando se aplica a los cuerpos; y las preguntas por la identidad de un cuerpo son muy a menudo preguntas sobre palabras. Pero cuando se aplica a las personas, la identidad no es ambigua y no admite grados, de más o de menos. Es el fundamento de todos los derechos y de todas las obligaciones, y de toda responsabilidad; y su noción es fija y precisa» [87].

Reid es evidentemente reduccionista por lo que hace a la identidad de los cuerpos. Según su opinión, la identidad personal es muy diferente. Implica un hecho que es siempre determinado, y que tiene que ser todo-o-nada. Es lo que llamo el hecho adicional profundo. Reid piensa que este hecho es el fundamento de la moral: que si, como he argumentado, no hubiera hecho semejante, no sería solamente cierto que no podemos ser «responsables» de delitos pasados, sino que se verían socavados todos los derechos y todas las obligaciones.

Algunos autores modernos hacen afirmaciones similares. Madell sostiene que «un análisis de la identidad personal en términos de continuidad psicológica... es totalmente destructivo de la gama completa de nuestras actitudes morales normales... La vergüenza, el remordimiento, el orgullo y la gratitud» dependen todos del rechazo de esta concepción [88]. Y Haksar afirma que mi modo de pensar socava todos los «derechos humanos» y todas las «restricciones morales no utilitaristas», y que es «incompatible con cualquier clase de moralidad humana» [89].

¿Deberíamos aceptar estas tesis radicales? Primero deberíamos notar lo siguiente. Si la verdad acerca de la identidad personal tuvie-

[87] En Perry (1), p. 112.
[88] Madell, p. 116.
[89] Haksar, p. 111.

ra estas implicaciones, la mayor parte de nosotros la encontraría profundamente perturbadora. Puede pensarse que, si estas *fuesen* las implicaciones de la Concepción Reduccionista, esto demostraría que es falsa. Pero no es así. La verdad puede ser perturbadora. Consideremos la afirmación de que el universo no fue creado por un Dios benevolente. Muchas personas la encuentran perturbadora: pero esto no puede demostrar que sea falsa. Si una verdad es perturbadora, esto no es razón para no creer en ella. Sólo puede ser una razón para actuar de ciertos modos. Podría ser una razón para tratar de ocultar esta verdad a otros. Podría hasta ser una razón para tratar de engañarnos a nosotros mismos, de manera que dejemos de creer en ella. Como he dicho, el pensamiento desiderativo es teóricamente irracional, pero puede ser racional desde el punto de vista práctico. Podría ser así, por ejemplo, si fuera el único medio de librarnos de la depresión profunda.

Consideremos a continuación una de estas tesis radicales. Hay gente que sostiene que, si es verdadera la Concepción Reduccionista, no podemos merecer ser castigados por nuestros crímenes. Este argumento asume que sólo el hecho adicional profundo lleva consigo merecimiento o responsabilidad. Haré de nuevo dos preguntas: (1) ¿Sería plausible o al menos justificable esta asunción si la Concepción No-Reduccionista fuera correcta? (2) ¿Es la asunción plausible, o justificable, dada la verdad de la Concepción Reduccionista?

Puede servir de ayuda volver a mi división imaginaria. Para la Concepción No-Reduccionista hay tres posibilidades, todas las cuales podrían ser la verdad. Podría ser verdadero que yo no vaya a ser ninguna de las dos personas resultantes, o que vaya a ser Izquierdo, o que vaya a ser Derecho. Supongamos que voy a ser Derecho. ¿Merecería ser castigado Izquierdo por los crímenes que yo cometí antes de la división? Un no-reduccionista podría responder de manera justificable:

> No. Izquierdo es tanto física como psicológicamente continuo contigo, tal y como tú eras antes de la división. Pero *él* no cometió tus crímenes. ¿Cómo va a merecer ser castigado por crímenes que cometió alguien distinto, en una época en la que él no existía? Sólo el hecho adicional profundo de la identidad personal lleva consigo responsabilidad por los crímenes pasados. En ausencia de este hecho, las dos continuidades no llevan consigo tal responsabilidad.

El no-reduccionista debería admitir que la continuidad psicológica tiene algunas implicaciones morales. Supongamos que mi crimen pasado hizo patente que soy un maníaco homicida. Como es psicológicamente continuo conmigo, Izquierdo también sería un maníaco homicida. Lo que podría justificar detenerle como medida preventiva aun antes de que cometa un crimen. Pero esto no demuestra que pueda merecer ser castigado por mis crímenes.

Supongamos ahora que el no-reduccionista cambia de opinión en lo que respecta a la naturaleza de la identidad personal. Si se hace reduccionista, puede afirmar de manera justificable que no merecemos castigo por nuestros crímenes. Según su concepción, el merecimiento requiere el hecho adicional profundo. Como no hay tal hecho, no hay merecimiento.

Como antes, aunque esta tesis radical sea defendible, puede ser negada de manera defendible. Puede servir de ayuda mencionar una analogía. Hay dos concepciones acerca del merecimiento y el determinismo. Según la concepción compatibilista, la clase de voluntad libre que se requiere para el merecimiento no sería socavada por la verdad del determinismo. Según la concepción incompatibilista, el determinismo socava tanto la voluntad libre como el merecimiento. Según esta segunda concepción, si fue causalmente inevitable que yo cometiera mi crimen, no puedo merecer ser castigado. Si está moralmente justificado mandarme a la cárcel, lo está sólo por razones utilitaristas. Una de tales razones es que mi encarcelamiento puede disuadir a otros de cometer crímenes. Y, evidentemente, sería irrelevante si se cree nada más que falsamente que he cometido un crimen: como incluso el culpable no merece castigo, no supondrá ninguna diferencia moral el que yo sea de hecho inocente.

Algunos afirman que sólo es defendible la concepción compatibilista. Otros dicen lo mismo de la incompatibilista. Un tercer grupo cree que este desacuerdo no ha sido resuelto de modo decisivo. Estos últimos podrían decir: «Aunque estas opiniones se con-

tradicen mutuamente, y por eso no pueden ser ambas verdaderas, ambas son defendibles. Nadie ha publicado todavía un argumento que refute decisivamente una opinión y establezca la otra».

Yo suscribo esta afirmación acerca de los diferentes pares de concepciones que he descrito. Los no-reduccionistas piensan que la identidad personal implica un hecho adicional profundo, distinto de la continuidad física y psicológica. Es una afirmación defendible que sólo este hecho lleva consigo el merecimiento por los delitos pasados y que, si no hay tal hecho, no hay merecimiento. Esta es la análoga de la concepción incompatibilista. Se puede mantener que el merecimiento es incompatible con el reduccionismo. Pero también se puede defender una concepción diferente: podemos afirmar de manera defendible que la continuidad psicológica lleva consigo merecimiento por los delitos pasados. Quizás haya un argumento que resuelva decisivamente este desacuerdo, pero yo todavía no lo he encontrado.

Consideremos a renglón seguido el hecho de que hay grados de conexividad psicológica. Supongamos que, entre un presidiario ahora y él mismo cuando cometió un crimen haya sólo conexiones psicológicas débiles. Esto usualmente será así sólo cuando alguien es condenado muchos años después de cometer el crimen. Pero podría ocurrir asimismo cuando hay una gran discontinuidad, como en la conversión de un joven italiano en busca de placeres en San Francisco. Podemos señalar la debilidad de las conexiones psicológicas llamando al presidiario el yo posterior del criminal.

No se verían afectadas dos razones para detenerle. Si un presidiario debiera ser reformado, o bien detenido preventivamente, esto depende de su estado presente y no de su relación con el criminal. Una tercera razón, la disuasión, depende de una pregunta diferente. ¿Se preocupan los criminales en potencia de tales yoes posteriores? ¿Se preocupan, por ejemplo, si no esperan ser cogidos durante muchos años? Si sí se preocupan, detener a sus yoes posteriores podría disuadir a otros.

¿Se merecería esto? Locke pensaba que si olvidamos nuestros delitos no merecemos castigo. Geach dice que esta opinión es

«moralmente repugnante» [90]. Y la mera pérdida de memoria sí que parece insuficiente. Los cambios de carácter son más relevantes. Pero el tema es complicado. Las afirmaciones sobre el merecimiento pueden apoyarse de manera verosímil con una gran variedad de argumentos. De acuerdo con algunos de ellos, la pérdida de memoria sería importante. Y de acuerdo con la mayoría, haría falta conocer la naturaleza y la causa de cualquier cambio de carácter.

No entraré en estos detalles. Pero haré una afirmación general. Cuando un presidiario está ahora menos íntimamente conectado a sí mismo en el momento del crimen, merece menos castigo. Si las conexiones son muy débiles, puede que no merezca ninguno. Esta afirmación parece plausible. Puede dar una de las razones por las que tenemos Estatutos de Limitaciones, que fijan períodos de tiempo después de los cuales no podemos ser castigados por nuestros delitos. (Supongamos que un hombre de noventa años, uno de los pocos poseedores legítimos del Premio Nobel de la Paz, confiesa que fue él quien, a la edad de veinte años, hirió a un policía en una pelea de borrachos. Aunque fue un delito grave, puede que ahora el hombre no merezca ser castigado.)

Esta afirmación debería distinguirse de la idea de responsabilidad disminuida. No apela a la enfermedad mental, sino que en vez de ello trata al yo posterior del malhechor como si fuera un cómplice cuerdo. Igual que lo que merece una persona corresponde al grado de su complicidad con un malhechor, así sus merecimientos ahora por alguna fechoría pasada corresponden al grado de conexividad psicológica entre ella misma ahora y ella misma cuando perpetró el delito.

Podemos tener la tentación de protestar, «Pero fue en igual medida *su* delito». Esto es cierto. Y esta verdad sería una buena objeción si no fuéramos reduccionistas. Pero para la Concepción Reduccionista esta verdad es demasiado trivial como para refutar mi afirmación sobre la responsabilidad reducida. Es como la afirmación, «Todo cómplice es en igual medida un cómplice». Tal afirmación no puede demostrar que la complicidad no tenga grados.

[90] En su *God and the Soul* [*Dios y el alma*], p. 4.

En esta sección he descrito tres concepciones. Según la Tesis Radical, como la Concepción Reduccionista es verdadera, nadie merece jamás ser castigado. Como antes, esta afirmación es defendible, pero también puede ser negada de manera defendible. He dicho también que el debilitamiento de las conexiones puede reducir la responsabilidad. Esta afirmación me parece más plausible que su negación.

110. Compromisos

Si nos volvemos a los compromisos, son de aplicación afirmaciones similares. Según la Tesis Radical, como la Concepción Reduccionista es verdadera, nunca podemos estar obligados por compromisos pasados. Esta afirmación es defendible, pero también lo es su negación. Y es plausible afirmar que el debilitamiento de las conexiones reduciría la fuerza de un compromiso.

Sería tedioso dar una defensa similar de estas conclusiones. Por eso me vuelvo a una cuestión que no tiene análogo en el caso del merecimiento. Cuando consideramos los compromisos, el hecho de la identidad personal entra dos veces. Tenemos que considerar la identidad tanto del que hace una promesa como de la persona a quien se la hace. El debilitamiento de la conexividad puede reducir la obligación del *que la hace*. Pero, en el caso de la persona que *recibió* la promesa, cualesquiera implicaciones de la Concepción Reduccionista podrían ser bloqueadas deliberadamente. Podríamos solicitar promesas de la forma: «Te ayudaré a ti y a todos tus yoes posteriores». Si las promesas que se me hacen adoptan esta forma, no puede sostenerse que más tarde serán socavadas por un cambio en mi carácter, o por cualquier otro debilitamiento, en el resto de mi vida, en conexividad psicológica.

Aquí hay una asimetría. Una formulación similar no puede obligar de manera tan obvia al que hace una promesa. Yo podría decir, «Yo, y todos mis yoes posteriores, te ayudaremos». Pero podría objetarse que puedo obligar o comprometer sólo a mi yo presente. Esta objeción tiene alguna fuerza, puesto que se parece a la afirmación

plausible de que sólo puedo obligarme o comprometerme a mí mismo. En contraste, nadie niega que yo pueda prometerte que ayudaré a otras personas, como por ejemplo a tus hijos. Por eso está claro que puedo prometerte ayudar a tus yoes posteriores.

Tal promesa puede hacerse especialmente vinculante. Supongamos que tú cambias mucho más que yo. Puedo entonces considerarme comprometido, no contigo, sino con tu yo anterior. Puedo pensar por consiguiente que tú no puedes renunciar a mi compromiso. Esto sería como el compromiso, contraído con alguien que ya está muerto, de ayudar a sus hijos. No podemos ser liberados de compromisos semejantes.

Un caso así sería raro. Pero, puesto que ilustra algunos otros puntos, vale la pena dar un ejemplo. Consideremos

> *El Ruso del Siglo Diecinueve.* Dentro de varios años, un joven ruso heredará vastas propiedades. Puesto que tiene ideales socialistas, tiene la intención, ahora, de dar la tierra a los campesinos. Pero sabe que con el tiempo sus ideales pueden evaporarse. Para ponerse en guardia ante esta posibilidad, hace dos cosas. Primero, firma un documento legal, que automáticamente entregará la tierra, y que sólo puede ser revocado con el consentimiento de su esposa. Entonces le dice a esta, «Prométeme que, si alguna vez cambio de opinión, y te pido que revoques el documento, no darás tu consentimiento». Y añade, «Considero mis ideales como esenciales para mí. Si los perdiera, quiero que pienses que dejo de existir. Quiero que consideres a tu marido entonces, no como yo, el hombre que te pide esta promesa, sino sólo como su corrompido yo posterior. Prométeme que tú no harías lo que él te pidiera».

Esta petición, que emplea el lenguaje de los yoes sucesivos, parece comprensible y también natural. Y si la esposa de este hombre hiciera la promesa, y en su edad madura él le pidiera que revocase el documento, ella podría de manera plausible considerarse a sí misma como no liberada de su compromiso. Podría parecerle a ella como si tuviera obligaciones con dos personas diferentes. Podría pensar que hacer lo que su marido le pide ahora sería traicionar al

joven que amaba y con el que se casó. Y podría considerar lo que ahora le dice su marido como incapaz de absolverle a ella de la deslealtad a ese hombre joven— de la deslealtad al yo anterior de su marido.

Puede parecer que un ejemplo como este no necesita la distinción entre yoes sucesivos. Supongamos que te pido que me prometas que nunca me vas a dar cigarrillos, aunque te los suplique. Tú puedes pensar que yo no puedo, suplicándote, liberarte por las buenas de este compromiso. Y al pensar esto no necesitas negar que soy yo con quien estás comprometido.

Esto es cierto. Pero la razón es que la adicción nubla el juicio. Ejemplos similares podrían involucrar un gran estrés o un gran dolor, o —como con Ulises, atado al mástil mientras cantaban las sirenas— una tentación extraordinaria. Cuando nada nubla el juicio de una persona, la mayoría de nosotros es de la creencia de que la persona con la que estamos comprometidos siempre nos puede liberar. Siempre puede, si está en sus cabales, renunciar a nuestro compromiso. Pensamos así, cualquiera que sea el compromiso. Según esta concepción, el contenido de un compromiso no puede prevenir que se renuncie a él. Aquí ocurre como con la autoridad. Supongamos que un general les dice a sus tropas, «Les ordeno atacar al amanecer, y no hacer caso de ninguna orden contraria posterior». Y más tarde dice, «No hagan caso de mi última orden, y retírense». A pesar del contenido de la primera orden, sería esta segunda la que las tropas deberían obedecer.

Volviendo a la pareja rusa. Los ideales del joven se esfuman, y en su edad madura le pide a su mujer que revoque el documento. Aunque ella le prometió negarse, él declara que ahora la libera de su compromiso. He descrito dos modos en que ella podría pensar que no está liberada. Podría entender que el cambio de opinión de su marido demuestra que él no puede hacer ahora juicios bien meditados. Pero podemos suponer que ella no tiene tal pensamiento. Podemos suponer también que ella comparte nuestro modo de pensar acerca de los compromisos. Si esto es así, ¿cómo puede creer que su marido no puede liberarla de su compromiso? Ella puede creer esto sólo si piensa que no es en cierto sentido *él* con quien está

comprometida. He descrito ese sentido. Ella puede considerar que la pérdida de ideales por parte del joven implica su sustitución por un yo posterior.

Este ejemplo ilustra una afirmación general. Podemos considerar algunos sucesos de la vida de una persona, en ciertos aspectos, como el nacimiento o la muerte. No en todos los aspectos, pero más allá de estos sucesos la persona tiene yoes anteriores o posteriores. Pero puede ser sólo uno de la serie de yoes el que sea objeto de alguna de nuestras emociones, y al que apliquemos alguno de nuestros principios.

El joven socialista ruso considera sus ideales como esenciales para su yo presente. Le pide a su esposa que prometa a este yo presente no actuar contra esos ideales. Y, según esta manera de pensar, ella nunca puede ser liberada de su compromiso. El yo con quien ella está comprometida, al tratar de liberarla, dejaría de existir.

Esta no es una cuestión legalista. Es en parte una verdad sobre las creencias y las emociones de esta mujer. Ella no ama a su marido de mediana edad sino al joven con el que se casó. Por eso piensa que es a este joven al que debe ser leal. Podemos amar a alguien que está muerto, y creer que estamos comprometidos con él. Y el objeto de ese amor y de ese compromiso puede ser, no alguien que está muerto, sino el yo anterior de una persona viva.

Puede objetarse que, mediante el procedimiento de distinguir yoes sucesivos convenientemente, podríamos librarnos injustamente de nuestros compromisos, o de nuestro justo merecido. No es así. Yo podría decir, «No fui yo el que robó el banco esta mañana, sino sólo mi yo pasado». Pero otros podrían contestar de modo plausible, «Fuiste tú». Como no hay criterios fijos, podemos elegir cuándo hablar de un nuevo yo. Pero tales elecciones pueden ser insinceras, y se puede saber que lo son. Y también pueden expresar sinceramente creencias —creencias que no son ellas mismas elegidas—. Esto ocurre con la mujer de mi ejemplo. Que el joven a quien amaba y con quien se casó, en cierto sentido ha dejado de existir —que su cínico marido de mediana edad es como mucho el yo posterior de este joven— estas afirmaciones le parece a ella que expresan más verdad que la simple afirmación, «son la misma persona».

De la misma manera que podemos describir más fielmente la historia de Rusia si la dividimos en las historias del Imperio y de la Unión Soviética, ella puede describir más fielmente la vida de su marido, y sus propias creencias y emociones, si divide su vida en la de dos yoes sucesivos [92].

III. La condición separada de las personas y la justicia distributiva

Somos personas diferentes, cada una con su propia vida que vivir. Esto es verdadero según todas las concepciones de la naturaleza de la identidad personal. Pero es una verdad más profunda para la Concepción No-Reduccionista. Si la aceptamos, podemos considerar esta verdad como uno de los hechos fundamentales que subyacen a todas las razones para actuar. Este hecho ha sido llamado la *condición separada de las personas.*

Sidgwick pensaba que este hecho es el fundamento de la teoría del Propio Interés acerca de la racionalidad. Si lo que es fundamental es que somos personas diferentes, cada una con su propia vida que vivir, esto da apoyo a la afirmación de que el fin último supremamente racional, para cada persona, es que su propia vida marche lo mejor posible. Sidgwick pensaba que hay otro fin último igualmente racional, que las cosas marchen, en su conjunto, lo mejor posible para todos. Muchos están de acuerdo con él en que este es el

[92] Nabokov, p. 64: «Ellos decían que lo único en el mundo que amaba este inglés era Rusia. Muchas personas no podían comprender por qué no se había quedado allí. La respuesta de Moon a preguntas de este tipo sería invariablemente, "Preguntad a Robertson" (el orientalista) "por qué no se quedó en Babilonia". Habría surgido la objeción, perfectamente razonable, de que Babilonia ya no existía. Moon asentiría con una pícara sonrisa silenciosa. Veía en la insurrección bolchevique una determinada finalidad claramente definida. Mientras que admitía de buena gana que, en un corto plazo de tiempo, tras las fases primitivas, podría desarrollarse una civilización en la "Unión Soviética", sin embargo sostenía que Rusia había concluido y era irrepetible».

fin último que nos da la moralidad. Y algunos aceptan la idea de Sidgwick de que, cuando la moralidad entra en conflicto con el propio interés, no hay respuesta a la pregunta de qué tenemos más razón para hacer. Cuando comparaba las razones morales con las interesadas, ninguna le parecía a Sidgwick tener más peso que la otra.

Sidgwick sostenía esta concepción porque pensaba que la condición separada de las personas es una profunda verdad. Pensaba que una apelación a esta verdad le da al teórico del Propio Interés una defensa suficiente contra las pretensiones de la moralidad. Y sugirió que, si adoptásemos una concepción diferente de la identidad personal, podríamos refutar la teoría del Propio Interés. Yo he afirmado que esto es cierto.

La teoría del Propio Interés le parecía a Sidgwick fundarse en la condición separada de las personas. Ahora someteré a discusión una afirmación similar acerca de la moralidad. Esta afirmación cuestiona la posición moral de Sidgwick. Este pensaba que había un principio moral último, el de Benevolencia Imparcial. Como aceptaba la Teoría Hedonista del Propio Interés, su principio de benevolencia adoptó una forma hedonista. Según este modo de ver, nuestro fin moral último es la mayor suma neta de felicidad sin dolor, o de «conciencia deseable» sin «conciencia indeseable». Los utilitaristas que rechazan el Hedonismo asumen que el fin último es la mayor suma neta de beneficios menos cargas. Según cualquiera de sus versiones, la Concepción Utilitarista es *impersonal* en el sentido siguiente. Todo lo que importa son las cantidades de felicidad y sufrimiento, o de beneficios y cargas. No supone ninguna diferencia moral cómo se distribuyen estas cantidades entre diferentes personas.

Muchos rechazan este modo de pensar. Podrían decir: «Uno de nuestros fines morales últimos puede ser el fin utilitarista. Pero tenemos como mínimo otro. La felicidad y el sufrimiento, o los beneficios y las cargas, deben ser justamente compartidos como entre diferentes personas. Además del Principio Utilitarista necesitamos principios de Justicia Distributiva. Un ejemplo es el Principio de Igualdad. Según este, es malo que algunas personas resulten menos favorecidas que otras sin ninguna falta por su parte».

El argumento a favor de la igualdad a menudo se afirma que está fundamentado en la condición separada de las personas. Una afirmación tal podría ser: «Como es una profunda verdad que vivimos vidas diferentes, es un fin moral último que, en la medida en que tenemos iguales méritos, las vidas de cada uno deberían marchar igualmente bien. Si esto es imposible, debería por lo menos ocurrir que las vidas de cada uno tengan una oportunidad igual de marchar bien» [93].

Si dejamos de creer en la Concepción No-Reduccionista, ¿qué implica esto para el Principio de Igualdad y otros principios distributivos? Mis principales afirmaciones serán estas. Este cambio de manera de pensar presta apoyo a tres argumentos sobre estos principios. Dos de ellos implican que deberíamos dar a estos principios un *alcance mayor*, lo que los haría más importantes. Pero también puede que seamos llevados a darles *menos peso*, lo que los haría menos importantes. Por ello tenemos que preguntarnos cuál sería el efecto *neto*.

112. Tres explicaciones de la concepción utilitarista

Antes de avanzar estos argumentos, mencionaré dos afirmaciones relacionadas. Pueden ser presentadas de la siguiente manera. Los utilitaristas rechazan los principios distributivos. Aspiran a la mayor suma neta de beneficios menos cargas, sea cual sea la distribución. Diré que ellos *maximizan*.

Cuando nuestros actos sólo pueden afectar a una persona, la mayoría de nosotros acepta la maximización. No pensamos que debamos dar a alguien menos días felices para ser más justos en el

[93] Nozick (2), p. 33, escribe: «Las restricciones morales colaterales que gravitan sobre lo que podemos hacer, afirmo, reflejan el hecho de nuestras existencias separadas. Reflejan el hecho de que ningún acto de imposición de equilibrio moral puede tener lugar entre nosotros: no hay ninguna superación en peso moral de una de nuestras vidas por otras tal que lleve a un bien *social* general mayor. No hay ningún sacrificio justificado de algunos de nosotros por los demás. Esta idea radical, a saber, la de que hay diferentes individuos con vidas separadas *y* que *por tanto* nadie puede sacrificarse por los otros...» (el segundo énfasis mío).

modo en que los extendemos por las partes de su vida. Hay, desde luego, argumentos a favor de la dispersión de los placeres. Permanecemos frescos y tenemos más cosas que aguardar con ilusión. Pero estos argumentos no cuentan contra la maximización; nos recuerdan cómo lograrla.

Cuando nuestros actos pueden afectar a varias personas diferentes, los utilitaristas hacen declaraciones similares. Admiten nuevos argumentos para distribuir los placeres, como el que apela a los efectos de la privación relativa, o a la utilidad marginal decreciente. Pero los utilitaristas tratan la igualdad como un mero medio, no como un fin aparte.

Como la actitud que tienen ante los conjuntos de vidas es como la nuestra ante las vidas individuales, los utilitaristas ignoran los límites entre las vidas. Podemos preguntar, «¿Por qué?».

Aquí van tres sugerencias:

(1) Su método de razonamiento moral les conduce a pasar por alto estos límites.
(2) Creen que los límites carecen de importancia porque piensan que los conjuntos de vidas son como las vidas individuales.
(3) Aceptan la Concepción Reduccionista de la identidad personal.

La sugerencia (1) la ha hecho Rawls [94]. La podemos resumir así. Muchos utilitaristas responden preguntas morales con el método llamado del *observador imparcial*. Cuando uno de tales utilitaristas se pregunta a sí mismo, como observador, qué sería correcto, o qué preferiría él desde una perspectiva imparcial, puede *identificarse* con todas las personas afectadas. Puede imaginar que él mismo *sería todas* estas diferentes personas. Esto le conducirá a ignorar el hecho de que están afectadas personas *diferentes*, y a ignorar así las pretensiones de justa distribución entre estas personas.

La sugerencia (2) ha sido hecha por Gauthier y otros [95]. Según ella, los utilitaristas tienen que asumir que el género huma-

[94] Rawls, p. 27.
[95] Gauthier (2), p. 126.

no es un superorganismo, o creer, como algunos hinduistas, en una *Alma del Mundo* individual. Si las sugerencias (1) y (2) fuesen verdaderas, explicarían la Concepción Utilitarista de maneras que acabarían con ella. Es evidentemente un error ignorar el hecho de que vivimos vidas diferentes. Y la humanidad no es un superorganismo.

Yo sugiero (3). Según esta sugerencia, los utilitaristas rechazan los principios distributivos porque creen en la Concepción Reduccionista. Si esta concepción apoya su rechazo, esta tercera explicación apoya antes que socava la concepción utilitarista.

En el caso de algunos utilitaristas, la sugerencia (1) puede ser correcta. Muchos utilitaristas consideran las cuestiones morales como si ellos fueran observadores imparciales. Algunos de ellos pueden ser, como dice Rawls, observadores *que se identifican*. Pero puede haber también observadores *distanciados*. Mientras que un observador que se identifica se imagina a sí mismo siendo *todas* las personas afectadas, y un rawlsiano se imagina a sí mismo siendo *una* de las personas afectadas, sin saber cuál, un observador distanciado se imagina a sí mismo no siendo *ninguna* de las personas afectadas.

Algunos utilitaristas han sido observadores imparciales distanciados. Ellos no pasan por alto la distinción entre personas. Y, como observa Rawls, parece haber pocas razones por las que los observadores distanciados debieran ser llevados a ignorar los principios de la justicia distributiva. Si nos aproximamos a la moralidad de este modo distanciado —si no nos pensamos a nosotros mismos como potencialmente implicados— puede que estemos en cierto modo más inclinados a rechazar estos principios, puesto que no temeríamos que nosotros mismos pudiésemos llegar a ser una de las personas que resultan menos favorecidas. Pero esta aproximación particular a las cuestiones morales no explica suficientemente por qué estos utilitaristas rechazan los principios distributivos.

¿Es correcta la sugerencia (2)? Como *explicación* de la Concepción Utilitarista, (2) es falsa. Algunos seguidores de Hegel creían que una nación era un superorganismo. Para citar a un autor, una

nación «es un ser vivo, como un individuo» [97]. Pero los utilitaristas ignoran los límites nacionales, y no creen que la humanidad sea un ser individual semejante.

La sugerencia (2) se entiende mejor no como explicación de la Concepción Utilitarista sino como objeción a esta concepción. La sugerencia puede ser que esta concepción no puede estar justificada a no ser que la humanidad sea un Superorganismo, y que, como esto es falso, los utilitaristas se equivocan al rechazar los principios distributivos.

Yo sugiero una explicación diferente. Según la sugerencia (3), los utilitaristas ignoran la distribución porque aceptan la Concepción Reduccionista. (3) es compatible con (1). Algunos utilitaristas pueden ser observadores que se identifican, y además aceptar la Concepción Reduccionista. Pero (3) entra en conflicto con (2).

Puede parecer que hay aquí un misterio. Según la sugerencia (2), los grupos de personas se comparan con personas individuales. Esto es el reverso de la Concepción Reduccionista, que compara la historia de una persona con la de una nación, o grupo de personas.

Como estas dos concepciones comparan a las naciones con las personas, ¿cómo pueden ser concepciones diferentes?

La respuesta es esta. Cuando consideramos a las naciones, la mayoría de nosotros es reduccionista. Pensamos que la existencia de una nación no implica nada más que la existencia de sus ciudadanos, que viven juntos en su territorio y actúan juntos de ciertos modos. En contraste, cuando consideramos a las personas, la mayoría de nosotros cree en la Concepción No-Reduccionista. Creemos que nuestra identidad tiene que ser determinada. Esto no puede ser cierto a menos que una persona sea una entidad que existe separadamente, distinta de su cerebro, su cuerpo y sus experiencias. La mayoría de nosotros es por tanto reduccionista en lo que respecta a las naciones pero no en lo que respecta a las personas. Es la diferencia entre estas opiniones comunes la que explica las dos compa-

[97] Espinas, citado en Perry (3), p. 402.

raciones. La afirmación de que X es como Y asume típicamente la concepción común de Y. Por tanto, diremos, «Las personas son como las naciones» si somos reduccionistas por lo que respecta a ambas. Si somos no reduccionistas por lo que respecta a ambas, diremos en vez de eso, «Las naciones son como las personas». La creencia en superorganismos puede ser una concepción no reduccionista de las personas.

113. Cambiando el alcance de un principio

Como los utilitaristas rechazan los principios distributivos, piensan que los límites de las vidas carecen de significación moral. Según su manera de ver, la condición separada de las personas puede ignorarse. He descrito tres explicaciones para esta opinión. Ahora argumentaré que, a pesar de algunas complicaciones, la mía es la mejor explicación.

Consideremos

> *La carga del niño.* Tenemos que decidir si le imponemos a un niño una privación. Si lo hacemos, esto
>
> (i) redundará en el propio beneficio del niño cuando sea adulto, o bien
>
> (ii) redundará en un beneficio similar de alguien distinto —por ejemplo, el hermano pequeño del niño.

¿Importa moralmente si es verdadera (i) o (ii)?

La mayoría de nosotros contestaría: «Sí. Si va a ser en pro del beneficio futuro del propio niño, al menos no puede haber injusticia». Podríamos añadir la afirmación general de que imponer cargas útiles está justificado con más probabilidad si estas cargas redundan en el propio bien de la persona.

Los utilitaristas aceptarían esta afirmación, pero la explicarían de modo diferente. En vez de decir que estas cargas no pueden ser injustas, dirían que, en general, son más fáciles de soportar.

Para bloquear esta respuesta, podemos suponer que nuestro niño es demasiado pequeño como para ser animado de esta forma.

Esto simplifica la disputa. Los utilitaristas dirían: «Que sea correcto imponerle esta carga a este niño depende sólo de lo grande que vaya a ser el beneficio. No depende de quién se vaya a beneficiar. No supondría ninguna diferencia moral que el beneficio le llegara no al niño mismo sino a alguien distinto». Los no utilitaristas contestarían: «Por el contrario, si le llegase al niño mismo, esto ayudaría a justificar la carga. Si le llegase a alguien distinto, eso sería injusto».

¿Las dos concepciones de la naturaleza de la identidad personal dan apoyo a bandos diferentes en esta disputa?

Parte de la respuesta está clara. Los no utilitaristas piensan que es un hecho moral importante que sea el mismo niño el que, en la edad adulta, se beneficie. Este hecho es más importante según la Concepción No-Reduccionista, porque según ella la identidad entre el niño y el adulto es más profunda en su naturaleza. Según el modo de pensar reduccionista, lo que está implicado en esta identidad es menos profundo, y se da, más allá de la adolescencia, en un grado reducido. Si somos reduccionistas podemos comparar el debilitamiento de las conexiones entre el niño y su yo adulto con la ausencia de conexiones entre personas diferentes. Daremos más peso al hecho de que, en este ejemplo, al niño no le preocupa lo que le ocurrirá a su yo adulto. Que vaya a ser *él* el que reciba el beneficio puede parecernos por tanto menos importante. Podríamos decir, «No será *él* el que se beneficie. Será sólo su yo adulto».

La Concepción No-Reduccionista apoya la respuesta no utilitarista. ¿Se sigue de esto que la Concepción Reduccionista apoya la afirmación utilitarista? No. Podríamos decir, «Igual que sería injusto que sea alguien distinto el que se beneficia, si no va a ser el niño sino sólo su yo adulto, esto también sería injusto».

La cuestión tiene carácter general. Si somos reduccionistas, consideraremos las toscas subdivisiones en el interior de las vidas como semejantes, de ciertas maneras, a las divisiones entre las vidas. Por eso podemos llegar a tratar igual dos clases de distribución: dentro de las vidas y entre las vidas. Pero hay dos modos de tratarlas igual. Podemos aplicar a ambas los principios distributivos, o a ninguna.

¿Cuál de estas cosas podríamos hacer? Distinguí dos modos en que puede cambiar nuestra concepción moral. Podemos dar a los

principios distributivos un alcance diferente, y un peso diferente. Si nos hacemos reduccionistas, podemos ser llevados a dar a estos principios un alcance *mayor*. Puesto que consideramos las subdivisiones dentro de las vidas como comparables, de ciertas maneras, a las divisiones entre las vidas, podemos aplicar los principios distributivos incluso dentro de las vidas, como en la afirmación que acabamos de hacer acerca de imponerle cargas a un niño. Ampliando el alcance de los principios distributivos, nos estaríamos separando más de la Concepción Utilitarista. En este respecto la Concepción Reduccionista va en contra más bien que a favor de la Concepción Utilitarista.

114. Cambiando el peso de un principio

Volvamos a continuación a la segunda explicación de la Concepción Utilitarista. Gauthier sugiere que suponer que deberíamos maximizar a favor de la humanidad «es suponer que la humanidad es una superpersona» [98].

Para comprender esta sugerencia primero deberíamos preguntar por qué, dentro de una vida individual, podemos ignorar los principios distributivos. ¿Por qué es aquí moralmente permisible simplemente maximizar? Podría pensarse que porque no es una cuestión moral lo que hagamos con nuestras propias vidas. Aunque esto fuese cierto, no podría ser la explicación. Pensamos que puede ser correcto maximizar dentro de la vida de alguien distinto. La medicina nos proporciona ejemplos. Pensamos que los médicos hacen bien en maximizar en nombre de sus pacientes sin conocimiento. Harían bien en elegir una operación que diera a sus pacientes una suma total de sufrimiento más pequeña, aunque este sufrimiento llegase todo dentro de un período determinado. No pensamos que esto sería una injusticia para esta persona durante este período.

Algunos afirman: «Somos libres de maximizar dentro de una vida sólo porque es *una* vida». Esta afirmación apoya el cargo de

[98] Gauthier (2), p. 126.

Gauthier contra los utilitaristas. Apoya la afirmación de que seríamos libres de maximizar sobre diferentes vidas sólo si fuesen partes de una vida individual.

Cuando se les presentara este argumento, los utilitaristas negarían su premisa. Podrían decir: «Lo que justifica la maximización no es la unidad de una vida. El sufrimiento es malo, y la felicidad es buena. Es mejor que haya menos de lo que es malo, y más de lo que es bueno. Esto basta para justificar la maximización. Como no es la unidad de una vida lo que, dentro de esta vida, justifica la maximización, esta puede ser justificada sobre diferentes vidas sin la asunción de que la humanidad es una superpersona».

Una conexión con la Concepción Reduccionista es esta. Es sobre su base, antes que sobre la base de la Concepción No-Reduccionista, que la premisa del argumento de Gauthier es negada con más verosimilitud. Si la unidad de una vida es menos profunda, será más convincente afirmar que esta unidad no es lo que justifica la maximización. Esta es una de las maneras en que la Concepción Reduccionista proporciona algún apoyo a la Concepción Utilitarista.

Ampliaré estas observaciones. Hay dos clases de distribución: dentro de las vidas y entre las vidas. Y hay dos modos de tratarlas igual. Podemos aplicar principios distributivos a las dos, o no aplicarlos a ninguna.

Los utilitaristas no los aplican a ninguna. Yo sugiero que esto puede ser, en parte, porque aceptan la Concepción Reduccionista. Una sugerencia incompatible es que aceptan la concepción opuesta, creyendo que la humanidad es una superpersona.

Mi sugerencia puede parecer evidentemente equivocada si pasamos por alto el hecho de que hay dos rutas que llevan al abandono de los principios distributivos. Podemos no darles ningún alcance, o en vez de eso no darles ningún peso.

Supongamos que asumimos que la única ruta es el cambio de alcance. Esto lo sugiere la afirmación de Rawls de que «el utilitarista extiende a la sociedad el principio de elección para un solo hombre» [99]. Aquí la asunción es que la ruta que lleva al Utilita-

[99] Rawls, p. 28 y p. 141.

rismo es un cambio en el alcance, no de los principios distributivos, sino de su correlativo: nuestra libertad de ignorar estos principios. Si asumimos que la única ruta es un cambio en alcance, puede parecer verdaderamente que los utilitaristas tienen o bien que asumir que cualquier grupo de personas es como una persona individual (la sugerencia de Gauthier), o bien, por lo menos, olvidar que no lo es (la sugerencia de Rawls).

Describiré la otra ruta. Los utilitaristas puede que no estén negando que los principios distributivos tengan alcance. Puede que estén negando que tengan peso. A esta negación tal vez le preste algún apoyo la Concepción Reduccionista.

Más exactamente, mi sugerencia es esta. La Concepción Reduccionista sí que apoya un cambio en el alcance de los principios distributivos. Apoya dar a estos principios *más* alcance, de forma que se apliquen incluso dentro de una vida individual. Esto es lo que afirmé en el caso de la carga del niño. Es más probable que sea el reduccionista el que considere esta relación del niño con su yo adulto como una relación con una persona diferente. Es por ello más probable que sea él quien afirme que es injusto imponerle cargas a este niño simplemente para beneficiar a su yo adulto. Es sobre la base de la Concepción No-Reduccionista que podemos responder más verosímilmente: «esto no puede ser injusto, puesto que seguirá siendo precisamente *él* el que más adelante se va a beneficiar». Como veremos después, hay otro argumento que, según la Concepción Reduccionista, apoya una ampliación mayor en el alcance de los principios distributivos. Pero aunque de estas dos formas apoye la Concepción Reduccionista la ampliación del alcance de los principios distributivos, también apoya el darles a estos principios menos peso. Y si no les damos *ningún* peso, no significará diferencia alguna el que les hayamos dado un mayor alcance. Así es como el efecto neto podría ser la Concepción Utilitarista.

Esta sugerencia difiere de las otras del modo siguiente. Rawls observa que la Concepción Utilitarista parece conllevar «la combinación de todas las personas en una» [100]. Nagel, de modo parecido,

[100] Rawls, p. 27.

afirma que el utilitarista «trata los deseos... de las distintas personas como si fueran los deseos... de una persona masiva» [101]. Y he citado una afirmación similar de Gauthier. Según mi sugerencia, la Concepción Utilitarista puede ser apoyada no por la combinación de las personas sino por su desintegración parcial. Puede basarse en la idea de que la vida de una persona está menos profundamente integrada de lo que la mayoría de nosotros supone. Los utilitaristas pueden tratar los beneficios y las cargas, no como si todos ellos llegaran dentro de la misma vida, sino como si no representara ninguna diferencia moral en dónde llegaran. Y esta creencia puede estar parcialmente apoyada por la idea de que la unidad de cada vida, y, por tanto, la diferencia entre las vidas, es en su naturaleza menos profunda.

Al ignorar los principios de distribución entre diferentes personas, la Concepción Utilitarista es *impersonal*. Rawls insinúa que «confunde la impersonalidad con la imparcialidad» [102]. Esto sería así si el modo en que los utilitaristas tratan de ser imparciales les llevara a pasar por alto la diferencia entre las personas. Y esto puede afirmarse de los utilitaristas cuyo método de razonamiento moral sí que tiene este efecto. Puede afirmarse de un observador imparcial *que se identifica*, y cuyo método de razonamiento le lleva a imaginar que será él mismo todas las personas afectadas. Pero pocos utilitaristas han razonado así. Y, según mi sugerencia, los utilitaristas no tienen por qué confundir impersonalidad con imparcialidad. Su imparcialidad puede ser en parte apoyada por la Concepción Reduccionista de la naturaleza de las personas. Como escribe Rawls, «el principio regulativo correcto para algo depende de la naturaleza de esa cosa» [103].

115. ¿Puede ser correcto gravar a alguien simplemente para beneficiar a alguien distinto?

Ahora desarrollaré mi sugerencia. Los utilitaristas creen que los beneficios y las cargas pueden pesarse libremente unos contra

[101] Nagel (1), p. 134.
[102] Rawls, p. 190.
[103] Rawls, p. 29.

otras, aunque lleguen a personas diferentes. Esto se niega con frecuencia.

Podemos en primer lugar distinguir dos clases de pesaje. La afirmación de que una cierta carga *tiene objetivamente más peso* que otra es la afirmación de que es mayor. La afirmación de que *tiene moralmente más peso* que otra es la afirmación de que debemos aligerarla incluso al precio de no ser capaces de aligerar la otra. Observaciones similares se aplican al pesaje de diferentes cargas, y al pesaje de cargas contra beneficios. Vale la pena explicar cómo un beneficio puede ser mayor que una carga, o pesar más que ella objetivamente. Esto sería lo más evidentemente verdadero si, cuando se ha ofrecido la elección de tener ambos o no tener ninguno, todo el mundo elegiría tener ambos. Todo el mundo pensaría aquí que vale la pena soportar esta carga para conseguir ese beneficio. Para que esta sea una buena prueba, las personas tienen que estar interesadas por igual en las partes de su vida en que recibirían estos beneficios y estas cargas. Como la mayoría de la gente se preocupa menos por el futuro más distante, la prueba se aplica mejor preguntando a la gente si elegirían soportar esta carga antes de recibir el beneficio. Si creen que valdría la pena hacerlo así, esto sugiere que, en su caso, este beneficio pesa más, objetivamente, que esta carga.

Ciertas personas afirman que una carga no puede ser superada objetivamente por otra, si llegan dentro de vidas diferentes. Dicen que esas comparaciones interpersonales no tienen sentido. Si yo pierdo mi dedo y tú la vida, no tiene sentido decir que tu pérdida puede ser mayor que la mía. Ignoraré aquí este modo de pensar.

Otros dicen que las cargas y los beneficios en vidas diferentes no pueden ser pesados *moralmente*. Consideraré una parte de esta afirmación. Es la de que la carga de alguien no puede ser moralmente superada por meros beneficios para alguien distinto. Digo *meros* beneficios, porque no se tiene la intención de que la declaración niegue que pueda ser correcto gravar a alguien para beneficiar a alguien distinto. Esto lo podría requerir la justicia distributiva. Podemos correctamente imponer cargas fiscales al rico para beneficiar al pobre. Lo que la declaración niega es que semejantes actos

puedan justificarse únicamente desde la razón utilitarista de que el beneficio sea mayor que la carga.

Esta afirmación a menudo adopta formas matizadas. Puede restringirse a grandes cargas, o hacerse para afirmar que, para superar la carga de una persona, el beneficio a otros tiene que ser mucho mayor. Discutiré aquí esta afirmación en su forma más simple, porque la mayor parte de mis observaciones podrían aplicarse a las otras formas. Rawls pone la afirmación como sigue: «El razonamiento que equilibra las ganancias y las pérdidas de personas diferentes... está excluido» [105]. A esta la llamo la *objeción al equilibrio*.

Esta objeción descansa en parte sobre una afirmación diferente, la de que la carga de alguien no puede ser *compensada* por beneficios a alguien distinto. A esta la llamo la *tesis de la compensación*. Con una matización, esta tesis es evidentemente verdadera: nuestras cargas pueden, en cierto sentido, ser compensadas por beneficios a los que amamos. Pero no pueden ser compensadas por beneficios a otras personas.

La tesis de la compensación no puede negarse. Si hacernos reduccionistas afectase a nuestro modo de pensar acerca de esa tesis, los efectos serían estos. Podríamos, en primer lugar, extender la tesis incluso dentro de las vidas individuales. Así, en el ejemplo que puse, podríamos afirmar que la carga del niño no puede ser compensada por los beneficios a su yo adulto. O podríamos decir que no puede haber aquí compensación *plena*. Lo cual podría apoyar la afirmación de que la carga del niño sería moralmente superada sólo si el beneficio a su yo adulto fuese *mucho* mayor. Estas afirmaciones serían como las de que, cuando las conexiones psicológicas se han reducido de manera notable, merecemos un castigo menor por las acciones de nuestros yoes anteriores, y estamos menos comprometidos por ellas. Estas afirmaciones tratan partes de una vida débilmente conectadas como, en ciertos aspectos o en cierto grado, vidas diferentes. Tales afirmaciones, por consiguiente, cambian el alcance de nuestros principios. Si pensamos que, entre algunas partes de la

[105] Rawls, p. 28. Omito las palabras «como si fueran una persona», puesto que estoy preguntando si este razonamiento tiene que conllevar esta asunción.

misma vida, puede haber o menos o ninguna compensación, estamos cambiando el alcance de la tesis de la compensación. Supuesto el contenido de la Concepción Reduccionista, este es un cambio de alcance en la dirección correcta.

A renglón seguido podríamos darle a esta tesis menos peso. Nuestra razón sería la que sugerí antes. La compensación presupone identidad personal. Según la Concepción Reduccionista, pensamos que el hecho de la identidad personal a través del tiempo es menos profundo o supone menos. Por eso podemos afirmar que este hecho tiene menos importancia moral. Como este hecho lo presupone la compensación, podemos decir que el hecho de la compensación es él mismo menos importante moralmente. Aunque no se pueda negar, a la tesis de la compensación le podemos dar de este modo menos peso. (Aquí va otro ejemplo de esta distinción. Que es injusto castigar al inocente no puede negarse. Pero podemos no darle ningún peso a esta afirmación. Nuestra incapacidad de negarla no nos fuerza a creer en el merecimiento. Si no creemos en el merecimiento, tal vez porque seamos deterministas, podemos decir: «aunque sea malo castigar al inocente, castigar al culpable es igual de malo».)

Volvamos ahora a la objeción al equilibrio. A diferencia de la compensación, el concepto de *mayor peso moral que* no presupone identidad personal. Se *puede* negar, por tanto, la objeción al equilibrio.

La negación se podría poner en estos términos: «nuestras cargas no pueden ser compensadas por meros beneficios a alguien distinto. Pero pueden ser moralmente superadas por tales beneficios. Hasta puede ser correcto dar los beneficios antes que aligerar las cargas. Las cargas son superadas moralmente por los beneficios si son objetivamente superadas por estos beneficios. Todo lo que se necesita es que los beneficios sean mayores que las cargas. No es importante, como tal, a quién llegan ambos».

Esta es la respuesta del utilitarista; sería su respuesta a los muchos argumentos en que la objeción al equilibrio parece no distinguirse de la tesis de la compensación. Así, Rawls usa la frase, «no puede ser justificado por, o ser compensado por» [106]. Y Perry escribe, «La felicidad de un millón de algún modo fracasa por completo a la hora de compensar o siquiera mitigar la tortura de uno» [107]. Perry parece igualar esta afirmación irrebatible con la objeción al equilibrio, y esto es un error.

La Concepción Reduccionista le da algún apoyo a la respuesta del utilitarista. La objeción al equilibrio descansa, en parte, en la tesis de la compensación. La Concepción Reduccionista apoya tanto la afirmación de que hay menos alcance para la compensación como la de que la compensación tiene menos peso moral. La compensación tiene menos alcance y menos peso que los que habría tenido si la Concepción No-Reduccionista hubiera sido verdadera. Como la compensación es, de estos dos modos, menos importante moralmente, hay menos apoyo para la objeción al equilibrio. Podemos afirmar, por tanto, que la respuesta del utilitarista es más plausible de lo que sería si la Concepción No-Reduccionista fuese verdadera. Pero esta afirmación no implica que tengamos que aceptar la Concepción Utilitarista. Por eso decimos sólo que le da *algún* apoyo a esta concepción.

Estas declaraciones pueden explicarse de modo diferente. Aun los que ponen objeciones al equilibrio piensan que puede estar justificado imponerle cargas a un niño en pro de su mayor beneficio propio en un período posterior de su vida. Su afirmación es que la carga de una persona, mientras que puede ser moralmente superada por los beneficios que se le hagan a ella, no puede ser superada *jamás* por meros beneficios a otras. Se acepta que es así aunque los beneficios sean mucho mayores que la carga. La afirmación, de este modo, da a los límites entre las vidas —o al hecho de la no identidad— una significación aplastante. Permite dentro de la misma vida lo que prohíbe totalmente sobre diferentes vidas.

Esta afirmación sería más plausible ateniéndonos a la Concepción No-Reduccionista. Como el hecho de la identidad se piensa aquí que es más profundo, el hecho de la no identidad podría más verosímilmente dar la impresión de tener tal importancia. Según este modo de ver, es una profunda verdad que todo lo que forma

[106] Rawls, p. 61.

[107] Perry (3), p. 671.

parte de la vida de una persona es en la misma medida su vida. Si estamos impresionados por esta verdad —por la unidad de cada vida— los límites entre las vidas parecerán más profundos. Esto apoya la afirmación de que, en el cálculo moral, no pueden cruzarse estos límites. Si apoyamos la Concepción Reduccionista, estaremos menos impresionados por esta verdad. Consideraremos menos profunda en su naturaleza la unidad de cada vida, y una cuestión de grado. Por eso podemos pensar que los límites entre las vidas son menos como los que se dan entre, digamos, las casillas de un tablero de ajedrez —los que dividen lo que es todo blanco de lo que es todo negro azabache— y más como los límites entre países diferentes. Pueden parecer entonces menos importantes moralmente.

Puede objetarse:

> El reduccionista afirma que las partes de cada vida están menos profundamente unificadas. Pero no afirma que haya más unidad entre vidas diferentes. Los límites entre las vidas son, según su modo de pensar, igual de profundos.

Podríamos responder por nuestra parte:

> Si una unidad es menos profunda, también lo es la correspondiente desunión. El hecho de que vivamos diferentes vidas es el hecho de que no somos la misma persona. Si el hecho de la identidad personal es menos profundo, también lo es el hecho de la no identidad. No hay aquí dos hechos diferentes, uno de los cuales es menos profundo según la Concepción Reduccionista, mientras que el otro permanece igual de profundo. Hay nada más que un hecho, y la negación de este hecho. La condición separada de las personas es la negación de que todos seamos la misma persona. Si el hecho de la identidad personal es menos profundo, también lo es la negación de este hecho.

116. Un argumento para darle menos peso al principio de igualdad

Volvamos ahora a un principio diferente, el de igual distribución como entre personas con los mismos merecimientos. La mayoría de

nosotros da al Principio de Igualdad sólo un cierto peso. Pensamos, por ejemplo, que la desigualdad puede estar justificada si produce una ganancia suficiente en la suma total de beneficios.

Ateniéndonos a este modo de ver las cosas, nosotros no rechazamos el Principio Utilitarista. Estamos de acuerdo en que todo aumento en la suma de beneficios tiene valor moral. Pero insistimos en que se le tiene que dar peso también al Principio de Igualdad. Aunque toda ganancia en bienestar importa, también importa *quién* gana. Ciertas distribuciones son, afirmamos, moralmente preferibles. Debemos dar cierta prioridad a ayudar a los que han salido peor parados sin que sea culpa suya. Y deberíamos tratar de aspirar a la igualdad.

Los utilitaristas responderían: «Estas afirmaciones son plausibles. Pero las políticas que recomiendan son las mismas que tienden a incrementar el bienestar total. Esta coincidencia sugiere que debemos cambiar nuestro modo de pensar acerca del estatus de estas afirmaciones [108]. No deberíamos considerarlas como controles que deben operar sobre nuestro fin moral último, sino como guías que nos llevan a él. Verdaderamente, deberíamos valorar la distribución igualitaria, pero el valor radica en sus efectos típicos».

Esta respuesta podría desarrollarse del modo siguiente. La mayoría de nosotros cree que una simple diferencia en cuándo ocurre algo, si no afecta a la naturaleza de lo que ocurre, no puede ser moralmente significante. Desde luego, ciertas respuestas a la pregunta «¿Cuándo?» son importantes. No podemos ignorar la posición temporal de los sucesos. Y es incluso plausible afirmar que, si estamos planeando cuándo dar o recibir beneficios, deberíamos aspirar a una distribución igual a lo largo del tiempo. Pero aspiramos a ella sólo a causa de sus efectos. No pensamos que la igualdad de beneficios en diferentes momentos sea, como tal, moralmente importante.

[108] Cf. Sidgwick (1), p. 425: «el argumento utilitarista no puede ser juzgado equitativamente a no ser que tomemos completamente en cuenta la fuerza acumulativa que deriva del carácter complejo de la coincidencia entre Utilitarismo y Sentido Común».

Los utilitaristas podrían decir: «si como tal no importa *cuándo* ocurre algo, ¿por qué va a importar *a quién* le ocurre? Las dos son meras diferencias en posición. Lo que es importante es la *naturaleza* de lo que ocurre. Cuando elegimos entre políticas sociales, sólo tenemos necesidad de preocuparnos por cómo de grandes vayan a ser los beneficios y las cargas. Dónde llegan estos, si en el espacio, o en el tiempo, o entre personas, es una cuestión que, en sí misma, carece de importancia».

Parte de la disputa radica, entonces, en esto. Los no utilitaristas entienden que la pregunta «¿Quién?» es totalmente distinta de la pregunta «¿Cuándo?». Si se les pregunta por la descripción más simple posible de los hechos moralmente relevantes, su descripción puede ser atemporal, pero tiene que ser personal. Podrían decir, por ejemplo: «un beneficio a esta persona, el mismo beneficio a otra, una carga igual de grande a la primera persona...». Los utilitaristas en cambio dirían simplemente, «Un beneficio, el mismo beneficio, una carga igual de grande...».

Hay argumentos muy diferentes a favor y en contra de estas dos posiciones. Pregunto: ¿hacernos reduccionistas daría apoyo a una de ellas?

Afirmo que lo daría. Partiendo de la Concepción Reduccionista, es más plausible comparar la pregunta «¿Quién?» con la pregunta «¿Cuándo?», y describir los datos morales de modo impersonal. Es más plausible de lo que lo sería si la Concepción No-Reduccionista fuese verdadera.

Volvamos a la comparación de Hume. La mayoría de nosotros cree que la existencia de una nación no implica más que la existencia de un determinado número de personas asociadas. No negamos la realidad de las naciones, lo que negamos es que sean reales de modo separado o independiente. Su existencia implica sólo la existencia de sus ciudadanos, viviendo juntos de ciertos modos, en su territorio.

Esta creencia da apoyo a ciertas afirmaciones morales. Si no hay nada más en una nación que sus ciudadanos, es menos plausible considerar a la nación en sí misma como un objeto primario de deberes, o un poseedor de derechos. Es más plausible concentrarse en los ciudadanos, y considerarles menos como ciudadanos y más como personas. Por eso podemos, según este modo de ver las cosas, pensar que la nacionalidad de una persona es menos importante moralmente.

Desde la Concepción Reduccionista mantenemos creencias similares. Creemos que la existencia de una persona no conlleva nada más que la ocurrencia de sucesos mentales y físicos interrelacionados. No negamos que las personas existan. Y consentimos en que no somos series de sucesos —que no somos pensamientos y acciones, sino pensadores y agentes—. Pero esto es verdadero sólo porque describimos nuestras vidas adscribiéndoles pensamientos y acciones a las personas. Como he sostenido, podríamos dar una descripción completa de nuestras vidas que fuese impersonal: que no afirmara que las personas existen. Negamos que seamos no sólo conceptualmente distintos de nuestros cuerpos, acciones y experiencias, sino también reales de una forma separada. Negamos que una persona sea una entidad cuya existencia sea separada de la existencia de su cerebro y de su cuerpo, y de la ocurrencia de sus experiencias. Y negamos que la existencia continua de una persona sea un hecho adicional profundo, que tenga que ser todo-o-nada, y que sea diferente de los hechos de la continuidad física y psicológica.

Estas creencias apoyan ciertas afirmaciones morales. Se hace más plausible, cuando pensamos en términos morales, concentrarse menos en la persona, el sujeto de las experiencias, y en vez de eso concentrarse más en las experiencias mismas. Se hace más plausible afirmar que, igual que hacemos bien al ignorar si las personas proceden de las mismas o diferentes naciones, hacemos bien al ignorar si las experiencias proceden del interior de la misma vida o de vidas diferentes.

Consideremos el alivio del sufrimiento. Supongamos que podemos ayudar sólo a una de dos personas. Conseguiremos más si ayudamos a la primera; pero es la segunda la que sufrió más en el pasado. Los que creen en la igualdad pueden decidir ayudar a la segunda persona. Lo que será menos efectivo, de modo que la cantidad de sufrimiento en las vidas de las dos personas será, en suma, más grande; pero

las cantidades en cada vida se harán más iguales. Si aceptamos la Concepción Reduccionista, podemos decidir de otra manera. Podemos decidir hacer lo máximo que podamos para aliviar el sufrimiento.

Para dar una idea de por qué, podemos variar el ejemplo. Supongamos que podemos ayudar sólo a una de dos naciones. A la que podemos ayudar más es una cuya historia en los siglos recientes fue más afortunada. La mayoría de nosotros no pensaría que pudiera ser correcto dejar que la humanidad sufriera más para que el sufrimiento estuviera más igualitariamente dividido entre las historias de las diferentes naciones. Al tratar de aliviar el sufrimiento, no consideramos a las naciones como las unidades moralmente significativas.

Con la Concepción Reduccionista comparamos las vidas de las personas con las historias de las naciones. Por eso podemos pensar lo mismo sobre ellas. Podemos pensar que, cuando estamos tratando de aliviar el sufrimiento, ni las personas ni las vidas son las unidades moralmente significativas. Podemos decidir de nuevo aspirar al mínimo sufrimiento posible, cualquiera que sea su distribución.

«El Utilitarismo», escribe Rawls, «no se toma en serio la distinción entre personas» [109]. Si «la condición separada de las per-

[109] Rawls, p. 27. Yo añado un comentario sobre otra observación. Se da, escribe Rawls:

> «... una curiosa anomalía... Es la costumbre pensar en el Utilitarismo como individualista, y ciertamente hay buenas razones para ello. Los utilitaristas...sostenían que el bien de la sociedad está constituido por las ventajas que disfrutan los individuos. Pero el Utilitarismo no es individualista... al aplicar a la sociedad el principio de elección por un hombre» (p. 29).

Mi consideración sugiere la explicación. Un individualista afirma (1) que el bienestar de una sociedad no consiste en otra cosa que en el bienestar de sus miembros, y (2) que los miembros tienen derecho a la parte que en justicia les corresponde. Supongamos que somos no reduccionistas acerca de las sociedades o de las naciones. Creemos que la existencia de una sociedad, o de una nación, trasciende la de sus miembros. Esta creencia amenaza la afirmación (1); en la busca de un fin nacional trascendente, la parte que en justicia nos corresponda puede parecer menos importante. Los no reduccionistas respecto de las naciones pueden así rechazar las dos afirmaciones del individualista. Los utilitaristas rechazan (2) pero aceptan (1). Esto sería, como afirma Rawls, «una curiosa anomalía», en caso de que la Concepción Utilitarista descansara sobre el No Reduccionismo respecto de

sonas... es *el* hecho básico de la moral» [110], esta es una acusación grave. He intentado demostrar que, si apelan a la verdad en lo que respecta a la naturaleza de las personas, los utilitaristas pueden ofrecer alguna defensa.

las naciones. Si este fuese su fundamento, esperaríamos que ellos rechazasen las dos afirmaciones. Yo he descrito un fundamento diferente. En vez de ser no reduccionistas respecto de las naciones, los utilitaristas pueden ser reduccionistas respecto de las personas. Esto suprimiría la anomalía. Los utilitaristas también son reduccionistas respecto de las naciones, y este doble Reduccionismo presta cierto apoyo a la Concepción Utilitarista. Si somos reduccionistas respecto de las naciones, entonces podemos aceptar con más plausibilidad la primera de las afirmaciones del individualista: que el bienestar de una nación no consiste en otra cosa más que en el de sus ciudadanos. Si también somos reduccionistas respecto de las personas, entonces podemos rechazar con más plausibilidad la segunda afirmación, la demanda de la parte que en justicia nos corresponde. Podemos dar menos peso a la unidad de cada vida y a la diferencia entre las vidas, y asignar más peso a las diversas experiencias que, juntas, constituyen estas vidas. Podemos decidir así que lo que es moralmente importante es sólo la naturaleza de lo que ocurre, no a quién ocurre. Podemos decidir que es siempre correcto incrementar los beneficios y reducir las cargas, cualquiera que sea su distribución. Cf. la observación en Anschutz de que «El principio del individualismo de Bentham», a diferencia del de Mill, «es completamente un principio de transición», puesto que «Bentham está diciendo que... como una comunidad es reducible a los individuos de los que se dice que son sus miembros, así también son los individuos reducibles, al menos para los propósitos de la moral y la legislación, a los placeres y dolores que decimos que sufren» (pp. 19-20). El Utilitarismo de Bentham puede venir apoyado en parte por su creencia en la Concepción Reduccionista. Pero mis afirmaciones no pueden aplicarse a todos los utilitaristas. La excepción obvia es Sidgwick. En pp. 416-17 de *The Methods* [*Los métodos*] Sidgwick da algún peso al Principio de Igualdad. Pero el Principio de Utilidad tiene prioridad absoluta. El rechazo por parte de Sidgwick de los principios distributivos no puede explicarse de la manera que he discutido. Sidgwick rechazaba la Concepción Reduccionista de Hume. Y terminó la primera edición de su libro con la palabra «fracaso» sobre todo porque asignaba tal peso a la distinción entre las personas. (Véase, por ejemplo, Sidgwick (1), p. 404, o la observación de la p. 498, «La distinción entre un individuo cualquiera y otro es real y fundamental».) Como he dicho, la condición separada de las personas le pareció a Sidgwick lo suficientemente profunda como para apoyar un rechazo por parte del teórico del Propio Interés de las afirmaciones de la moralidad. Me habría gustado poder preguntarle por qué este hecho profundo, a su modo de ver, no apoyaba las demandas de la justa distribución.

[110] Findlay, p. 294.

Como he declarado, no puede ser una defensa completa. La cuestión es si deberíamos aceptar los principios de la justicia distributiva, y, en caso afirmativo, cuánto peso debemos dar a estos principios. Yo afirmo que, según la Concepción Reduccionista, es más plausible darles menos peso, e incluso ningún peso en absoluto. Pero esto sólo significa, «más plausible de lo que lo sería según la Concepción No-Reduccionista». Es compatible con mi afirmación el que, incluso según la concepción reduccionista, sea inverosímil no darles a estos principios ningún peso.

El argumento que he dado no es capaz de demostrar que tengamos que aceptar la Concepción Utilitarista. Sólo puede brindarnos conclusiones sobre la plausibilidad relativa. Cuando dejamos de creer que las personas son entidades que existen separadamente, la Concepción Utilitarista se hace más plausible. ¿Es la ganancia en plausibilidad grande o pequeña? Mi argumento deja abierta esta pregunta.

117. Un argumento más radical

Finalizo con otro argumento. Cuando dejamos de ser no reduccionistas, no creemos simplemente que la identidad personal sea menos profunda, o implique menos. Vamos a parar a la creencia de que no implica el hecho adicional profundo.

Volvamos a la pregunta de qué justifica maximizar dentro de una vida. ¿Por qué puede ser correcto imponerles a los niños grandes cargas, para que reciban beneficios aun mayores cuando sean adultos? Algunos piensan que esto está justificado sólo por la unidad de la vida de estas personas. Cuando los niños sean mayores, sus cargas serán completamente compensadas por estos beneficios mayores.

Podría hacerse una afirmación diferente. Podría decirse que lo que justifica la maximización es el hecho adicional profundo. Sólo este hecho hace posible que un beneficio posterior compense una carga anterior. La compensación no sólo presupone identidad personal. Según este parecer, presupone el hecho adicional profundo [111].

[111] Wachsberg defiende este modo de ver las cosas. Wachsberg añade y convincentemente defiende la afirmación de que la creencia en el Hecho Adicional conlleva la idea de que todas las experiencias ocurren en «la misma conciencia», cuya mismidad a través del tiempo es una mala generalización de la unidad de la conciencia en un momento determinado.

Volvamos al caso en que me divido. Asumamos que creemos en la concepción no reduccionista, y que suponemos que yo seré Derecho. Antes de la división, yo tenía más que la parte que en justicia me corresponde de muchos recursos, viviendo en la opulencia durante muchos años. Tras la división, Izquierdo y yo obtendremos cada uno menos de la parte que en justicia nos corresponde. Se afirma que esto está justificado, en mi caso, porque tendrá el resultado de que en mi vida, considerada en su conjunto, yo recibiré la parte que me corresponde. Mi parte menor de ahora fue compensada enteramente de antemano por mi parte mayor antes de la división.

¿Podríamos decir lo mismo, verosímilmente, de Izquierdo? ¿La continuidad psicológica hace posible la compensación, aun en ausencia de la identidad personal? Es defendible responder:

> No. Izquierdo nunca disfrutó una parte más grande. *Él* no disfrutó estos años de opulencia. Es irrelevante que pueda cuasi-recordar *tu* disfrute de esta opulencia. Es irrelevante que sea física y psicológicamente continuo con alguien que tuvo más que la parte que en justicia le correspondía, en una época en que Izquierdo no existía. Ahora sería injusto darle a Izquierdo menos que la parte que le corresponde. En ausencia de la identidad personal, la continuidad psicológica no puede hacer posible la compensación.

Supongamos a renglón seguido que llegamos a creer en la Concepción Reduccionista. Habíamos afirmado, de manera justificable, que sólo el hecho adicional profundo hace posible la compensación. Habíamos afirmado que, como demuestra el caso de Izquierdo, la continuidad física y psicológica no puede por sí misma hacer posible la compensación. Ahora creemos que no hay hecho adicional profundo, y que la identidad personal sólo consiste en estas dos clases de continuidad. Como podríamos afirmar de manera justificable que sólo este hecho adicional profundo hace la compensación posible, y como no hay semejante hecho, podemos con-

cluir de manera justificable que no puede haber compensación a través del tiempo. Podemos afirmar que un beneficio en una época no puede proporcionar compensación para una carga en otra, aunque ambos lleguen dentro de la misma vida. Sólo puede haber compensación simultánea, como cuando el dolor de exponer mi rostro al viento helado es completamente compensado por la visión del panorama sublime desde la montaña que he escalado [112].

Como esta conclusión es defendible, este argumento es más poderoso que los que apoyan la Concepción Utilitarista. Esos argumentos no demostraron que esta concepción sea defendible. Pero, como en el caso de los argumentos anteriores que apelaban al hecho adicional, aunque esta nueva conclusión sea defendible, puede ser también negada de forma defendible.

Esta nueva conclusión lleva consigo otro cambio en el alcance de nuestros principios distributivos. La aplicación de estos principios depende del alcance de la posible compensación. Según el argumento que se acaba de dar, no puede haber compensación a través del tiempo, ni siquiera dentro de la misma vida. Consideremos el Principio de Igualdad. Según el argumento que se acaba de dar, no deberíamos aspirar a la igualdad ni entre vidas diferentes ni entre las partes diferentes de la misma vida. Deberíamos aspirar a la igualdad entre los estados en los que están las personas en momentos particulares.

Nagel hace unas observaciones relevantes:

> «Nótese que estos pensamientos no *dependen* de ninguna idea concreta de la identidad personal a través del tiempo, aunque puedan *emplear* tal idea. Todo lo que se necesita para evocarlos es una distinción entre personas en un momento dado. El impulso a la igualdad distributiva surge mientras podamos distinguir entre dos experiencias que son tenidas por dos personas y dos experiencias que son tenidas por una persona. Los criterios de identidad personal a través del tiempo simplemente determinan el tamaño de las unidades sobre las que opera un principio distributivo. Eso, para decirlo en pocas palabras, es lo que pienso que está equivocado en la explicación que da Parfit de la relación entre justicia distributiva e identidad personal» [113].

En la explicación que discute Nagel, yo distinguí entre cambiar el alcance de nuestros principios morales, y dar a estos principios un peso diferente. Y afirmé que, si nos desplazamos de la Concepción No-Reduccionista a la Reduccionista, podemos dar a estos principios, con más verosimilitud, un mayor alcance pero menos peso [114]. Nagel dice que el efecto sólo puede ser dar a estos principios mayor alcance —que yo me equivoco al afirmar que un cambio de concepción puede afectar también al peso que damos a estos principios—. Pero, ¿qué es lo que hay de malo en esta afirmación? ¿Por qué debería ser el efecto sólo relativo al alcance? Un cambio de concepción sobre los hechos a menudo hace plausible dar a un principio moral un peso diferente. Si esto no puede ser plausible en el caso presente, hace falta demostrarlo. Creo que no se podría.

El argumento que ahora estoy discutiendo apoya la clase de efecto que Nagel aprueba. Apela a un cambio de opinión en lo que respecta al criterio de identidad personal. Al hacernos reduccionis-

[112] L. Temkin sostiene (en correspondencia) que si un reduccionista niega la compensación a través del tiempo, también debería negar la compensación simultánea. La unidad de la conciencia en un momento determinado no es un hecho más profundo que la continuidad psicológica a través del tiempo. Parece que esto se pone de manifiesto cuando consideramos nuestros recuerdos a corto plazo de los últimos pocos momentos, o del *presente especioso*. No puede haber una profunda diferencia entre esta continuidad a corto plazo y la unidad de la conciencia en un momento determinado. Lo cual presta apoyo a la sugerencia, hecha abajo, de que sobre la base de esta concepción nuestros principios distributivos coincidirían aproximadamente con el Utilitarismo Negativo. Si puede haber compensación simultánea, como en mi caso del viento helado y la vista sublime, el peor conjunto de experiencias en un momento temporal singular no tienen por qué ser las que conlleven el dolor o el sufrimiento más intensos, puesto que este dolor podría ser más que compensado por alguna otra experiencia en este conjunto. Si no puede haber siquiera compensación simultánea, nuestro interés consistiría en mejorar no el peor conjunto de experiencias, sino los peores miembros de tales conjuntos. Y es más probable que estos sean dolores intensos, u otras clases de sufrimiento extremo.

[113] Nagel (4), pp. 124-5, nota a pie de página 16.
[114] Parfit (3), p. 153.

tas, dejamos de pensar que la identidad personal implica el hecho adicional profundo. El argumento afirma que lo que la compensación presupone no es la identidad personal según cualquier concepción, sino la identidad personal según la Concepción No-Reduccionista. La compensación presupone el hecho adicional profundo. La continuidad psicológica, en ausencia de este hecho, no puede hacer posible la compensación a través del tiempo.

Si esto no es posible, ¿qué nos dirán que hagamos nuestros principios distributivos? Coincidirán aproximadamente con el *Utilitarismo Negativo*: la concepción que da prioridad al alivio del sufrimiento. Nagel habla de la *unidad* sobre la cual opera un principio distributivo. Si esta unidad es la totalidad de la vida de una persona, como asumen Rawls y muchos otros, un Principio de Igualdad nos dirá que tratemos de ayudar a las personas que están peor. Si la unidad es el estado de cualquier persona en un momento determinado, un Principio de Igualdad nos dirá que tratemos de mejorar, no la vida de las personas que están peor, sino los peores estados en los que las personas se encuentran.

Los igualitarios pueden no estar de acuerdo en el aspecto en el que deberíamos tratar de hacer iguales a las personas. ¿Deberíamos aspirar a la igualdad de bienestar, o a la igualdad de recursos? Supongamos que una vez creíamos que las unidades relevantes para los principios distributivos son las vidas completas, y que ahora creemos que las unidades relevantes son los estados en los que se encuentran las personas en momentos determinados. Este cambio de nuestra opinión hace menos plausible afirmar que nuestra preocupación debería ser la igualdad de recursos. Alguien con muchísimos recursos puede encontrarse en un estado muy malo en un momento determinado, y sus vastos recursos tal vez no le sirvan para aliviar este estado. Si las unidades relevantes son los estados en los que se encuentran las personas en momentos determinados, es más plausible afirmar que lo que nos debería preocupar es la calidad de las experiencias de las personas en esos momentos. Lo que corresponderá a los que están peor, según la idea de Rawls, serán las experiencias particulares que son peores, o las más indeseables. Estas serán las experiencias de gran dolor físico, o de gran angustia

o aflicción. Nuestro Principio de Igualdad nos dirá que tratemos de prevenir o de mejorar estas experiencias, y darle a esto prioridad sobre la promoción de experiencias deseables.

Cuando cambiamos la unidad desde la vida completa a las experiencias de las personas en momentos determinados, ¿deberíamos dar un peso diferente a nuestros principios distributivos? Por lo menos un autor dice que sí. Wachsberg sostiene que, si cambiamos el alcance de nuestros principios distributivos de este modo, no deberíamos dar a estos principios *ningún* peso. Si las unidades relevantes son las experiencias que las personas tienen en un momento determinado, Wachsberg cree que los principios distributivos pierden su verosimilitud [115]. Yo pienso, más precavidamente, que se hacen menos plausibles. Me parece plausible que el alivio del sufrimiento tenga mayor peso que el que le dan los utilitaristas corrientes o positivos. Pero no parece plausible que deba tener una prioridad absoluta, ni siquiera mucho mayor peso.

Haksar apunta una razón para pensar que, según la Concepción Reduccionista, los principios distributivos deberían tener un peso menor. Escribe:

> «... si la teoría de Parfit es correcta, si no hay individuos persistentes (excepto en un sentido trivial), ¿por qué deberíamos preocuparnos tanto por el sufrimiento del mundo? El sufrimiento todavía sería real, pero cuánto peor es cuando (intrínsecamente) el mismísimo individuo sigue sufriendo una y otra vez» [116].

Según el argumento que ahora estoy discutiendo, no puede haber compensación a través del tiempo sin el hecho adicional profundo. Como no existe semejante hecho, no hay compensación semejante. Si hacemos esta afirmación, no podemos afirmar también que una cierta cantidad de sufrimiento será un mal peor si, en vez de dispersarse por diferentes vidas, se concentra y se prolonga en la vida de una persona determinada. La afirmación de que esto es

[115] *Op. cit.*
[116] Haksar, p. 111.

un mal peor asume que la unidad relevante se extiende a lo largo del tiempo, lo cual niega este argumento. Si la maldad del sufrimiento es sólo la maldad que tiene en momentos determinados, la afirmación de Haksar parece plausible. Según este modo de pensar, aunque concentrado y prolongado el sufrimiento sea malo, no es tan malo como lo sería si la Concepción No-Reduccionista fuese verdadera. Esto reduce la plausibilidad del Utilitarismo Negativo.

Estas observaciones pueden malentenderse con facilidad. Según cualquier opinión acerca de la identidad personal, o acerca del alcance de los principios distributivos, el sufrimiento de una persona será de hecho mayor y más duro de soportar, si esta persona sabe que se va a prolongar. Esto proporciona una poderosa razón, hasta a los ojos de los utilitaristas positivos, para dar prioridad a tratar de prevenirlo. Mis observaciones han tratado sobre si tenemos una razón moral *adicional* para dar prioridad a la prevención del sufrimiento, sea o no sea prolongado. Que tenemos tal razón es lo que el utilitarista negativo afirma. Y hemos encontrado que, si cambiamos el alcance de nuestro Principio de Igualdad, de manera que las unidades relevantes sean los estados en que se encuentran las personas en momentos determinados, nuestro Principio de Igualdad coincide aproximadamente con el Principio de Utilidad Negativa. ¿Qué peso le debemos dar a estos dos principios? A diferencia de Wachsberg, creo que estos principios siguen siendo plausibles. Pero creo que tienen menos plausibilidad que la que tenían cuando creíamos que la unidad relevante era el todo de la vida de una persona. Y esto es así en parte por la razón de Haksar. Si la unidad de la vida de una persona no implica el hecho adicional profundo, es una afirmación defendible que no puede haber compensación a través del tiempo. Y si este tipo de unidad no puede hacer posible la compensación, no puede hacer posible el mal total que sería el sufrimiento prolongado si fuese verdadera la Concepción No-Reduccionista.

118. Conclusiones

Resumiré ahora las principales afirmaciones de esta larga discusión. Pregunté cómo podría afectar a nuestras emociones, y a nuestras

creencias sobre la racionalidad y la moralidad, un cambio de opinión en lo que respecta a la naturaleza de la identidad personal. Y acabo de discutir cómo, si cambiamos de opinión, esto podría afectar a nuestras creencias sobre el Principio de Igualdad y los otros principios distributivos. Antes he argumentado que, si vamos de la Concepción No-Reduccionista a la Reduccionista, se hace más plausible que hay menor alcance para la compensación dentro de la misma vida. De modo que es más plausible afirmar que grandes cargas impuestas a un niño no pueden compensarse, o compensarse del todo, por beneficios de algún modo mayores en su vida adulta. Cuando ampliamos así los principios distributivos para que cubran tanto vidas completas como partes de la misma vida débilmente conectadas, esto los hace más importantes. Lo que significa un distanciamiento de la Concepción Utilitarista.

Acabo de discutir un segundo argumento a favor de un cambio en el alcance de nuestros principios distributivos. Afirma que sólo el hecho adicional profundo hace posible la compensación a través de las diferentes partes de una vida. Como tal hecho no existe, debemos cambiar, como apunta Nagel, «el tamaño de las unidades sobre las que opera un principio distributivo». Las unidades se reducen a los estados de las personas en momentos determinados. Esta conclusión se puede defender, pero también su negación.

Dado este cambio en su alcance, el Principio de Igualdad viene a coincidir aproximadamente con el Utilitarismo Negativo. Si aplicamos este principio a los estados de las personas en momentos determinados, ¿deberíamos darle tanto peso? Sugiero que no.

Un tercer argumento afirmaba que, sea cual sea su alcance, deberíamos dar menos peso a los principios distributivos. A menudo se entiende que estos se fundan en la condición separada, o no identidad, de las diferentes personas. Este hecho es menos profundo desde la Concepción Reduccionista, puesto que la identidad es menos profunda. No implica el hecho adicional en el que estamos inclinados a creer. Como el hecho sobre el que están fundados se ve que es menos profundo, es más plausible darles un peso menor a los principios distributivos. Si dejamos de creer que las personas son entidades que existen separadamente, y llegamos a creer que la uni-

dad de una vida no implica más que las diversas relaciones que se dan entre las experiencias de esta vida, se hace más plausible preocuparse más por la calidad de las experiencias, y menos por de quién son. Esto da algún apoyo a la Concepción Utilitarista, haciéndola más plausible de lo que lo hubiera sido si la Concepción No-Reduccionista hubiera sido verdadera. Aunque no nos demos cuenta de ello, la mayoría de nosotros se inclina a creer en la Concepción No-Reduccionista. La impersonalidad del Utilitarismo es por eso menos inverosímil de lo que la mayor parte de nosotros cree.

Antes discutí si, en caso de cambiar de opinión, seguíamos teniendo las mismas razones para estar especialmente preocupados por nuestro propio futuro. Algunos autores dicen que sólo el hecho adicional profundo justifica esta preocupación especial, y que, como no existe semejante hecho, no tenemos razones para estar especialmente preocupados por nuestro propio futuro. Esta tesis radical es defendible, pero también lo es su negación.

Entonces avancé otro argumento contra la teoría del Propio Interés sobre la racionalidad. Apelaba al hecho de que parte de lo que es importante en la identidad personal, la conexividad psicológica, se da a lo largo del tiempo en grados reducidos. Cuando un hecho importante se da en un grado reducido, no puede ser irracional creer que este hecho tiene menos importancia. Por eso no puede ser irracional estar menos preocupado, ahora, por esas partes de nuestro futuro con las que ahora estamos menos estrechamente conectados. Supuesta la verdad de la Concepción Reduccionista, esto refuta la Teoría Clásica del Propio Interés.

Según la Teoría Revisada del Propio Interés, que no está refutada, puede no ser irracional hacer lo que se sabe que será peor para uno mismo. Puede no ser irracional la imprudencia grave. Si tales actos no son irracionales, lo cierto es que necesitan ser criticados. Afirmé que deberíamos considerarlos moralmente incorrectos. Si tales actos son moralmente incorrectos, esto refuerza la causa del paternalismo.

Si nos hacemos reduccionistas, podemos afirmar verosímilmente que un óvulo fertilizado no es un ser humano, y que llega a ser un ser humano sólo gradualmente durante el embarazo. Esto apoya

la tesis de que el aborto no es incorrecto en las primeras semanas, y que sólo gradualmente llega a serlo.

También consideré lo que, según la Concepción Reduccionista, deberíamos pensar sobre el merecimiento y los compromisos. Algunos autores afirman que sólo el hecho adicional profundo lleva consigo merecimiento, y que, como tal hecho no existe, no podemos merecer ser castigados por delitos pasados. Esta conclusión pareció defendible, pero asimismo su negación. Entonces argüí a favor de la afirmación general de que, si son débiles las conexiones entre un criminal ahora y él mismo en el momento de su crimen, merece un castigo menor. Afirmaciones similares se aplicarían al compromiso.

Debemos ser reduccionistas. Si esto es un cambio de concepción, apoya diversos cambios en nuestras creencias sobre la racionalidad y la moralidad. Hay otros cambios que no he discutido [117].

El efecto sobre nuestras emociones puede ser diferente para diferentes personas. Para los que aceptan las Tesis Radicales, el efecto puede ser perturbador. Describí el efecto que tienen sobre mí. Como yo niego estas tesis, encuentro la verdad liberadora y consoladora. Me hace preocuparme menos por mi propio futuro y por mi muerte, y más por los demás. Doy la bienvenida a esta ampliación de mi interés.

[117] No he discutido los efectos en las creencias que no tratan sobre la racionalidad y la moralidad. Uno de tales efectos puede ser reducir el *problema de las otras mentes*. Cuando yo era no reduccionista, creía que la unidad de mi vida no consistía meramente en las diversas conexiones que se dan entre mis experiencias, y los estados y procesos de mi cerebro. Creía que era un hecho adicional y fundamental el que todas estas experiencias fuesen mías. Lo cual me hacía más difícil imaginar lo que sería para algunas experiencias *no* ser mías, añadiendo importancia al Problema de las Otras Mentes. Según la Concepción Reduccionista, esta dificultad desaparece. Y la posibilidad de cuasi-recordar las experiencias pasadas de otras personas también puede ayudar a resolver este problema, como puede ayudar el caso imaginario en que me divido. Es una cuestión vacía la de cuál de las personas resultantes sería yo. (Como ya expliqué, esta es una cuestión vacía aunque haya una respuesta que sea la mejor descripción.) Como sería una cuestión vacía la de si los recuerdos aparentes de estas personas son meros cuasi-recuerdos de los contenidos de otra mente, estas personas podrían no tomarse en serio el Problema de las Otras Mentes.

Cuarta parte

LAS GENERACIONES FUTURAS

16

EL PROBLEMA DE LA NO-IDENTIDAD

Hay otra pregunta sobre la identidad personal. Cada uno de nosotros podría no haber existido jamás. ¿Qué habría hecho que esto ocurriese? La respuesta produce un problema que la mayoría de nosotros pasa por alto. 607

Uno de mis objetivos en la Cuarta Parte es discutir este problema. Mi otro objetivo es discutir la parte de nuestra teoría moral en que surge este problema. Es la parte que incluye cómo afectamos a las generaciones futuras, la más importante de nuestra teoría moral, desde el momento en que los próximos siglos serán los más importantes de la historia humana.

119. Cómo nuestra identidad depende de hecho de cuándo fuimos concebidos

¿Qué habría determinado que una persona particular jamás hubiera existido? Con una matización, yo creo en

> *La Tesis de la Dependencia Temporal*: Si una persona particular no hubiese sido concebida cuando de hecho fue concebida, es *de hecho* verdadero que nunca habría existido.

Esta tesis no es obviamente verdadera. Por eso una mujer escribe:

> «Es siempre fascinante especular con la idea de quiénes habríamos sido si nuestros padres se hubiesen casado con otras personas» [1].

Al preguntarse quién habría sido ella, esta mujer ignora la respuesta: «Nadie».

Aunque la tesis de la dependencia temporal no sea obviamente verdadera, no es polémica, y es fácil creer en ella. De modo que no es como la Concepción Reduccionista de la identidad personal a través del tiempo. Esta es una de varias concepciones rivales, y es difícil de creer. La Tesis de la Dependencia Temporal no trata acerca de la identidad a través del tiempo. Trata de un tema diferente aunque relacionado: la identidad personal en diferentes historias posibles del mundo. Vale la pena discutir diversos pareceres acerca de este tema. Pero la tesis de la dependencia temporal *no* es uno de ellos. Es una tesis que es verdadera para *todos* estos pareceres.

Como he dicho, la tesis debería ser matizada. Cada uno de nosotros surgió de un par de células concreto: un óvulo y el espermatozoide que, uno de entre millones, lo fecundó. Supongamos que mi madre no hubiese concebido un hijo en el momento en que de hecho me concibió a mí. Y supongamos que hubiese concebido un hijo dentro de unos pocos días alrededor ese momento. Este niño se habría originado del mismo óvulo concreto del que yo me originé. Pero aunque hubiese sido concebido sólo unos cuantos segundos antes o después, es casi seguro que se habría originado de un espermatozoide diferente. Este niño habría tenido algunos de mis genes, pero no todos. ¿Habría sido yo?

Estamos inclinados a creer que cualquier pregunta sobre nuestra identidad tiene que tener una respuesta, que tiene que ser Sí o No. Como antes, yo rechazo este modo de pensar. Hay casos en los que nuestra identidad es indeterminada. El que acabo de describir puede ser un caso así. Si lo es, mi pregunta no tiene respuesta. No es ni verdadero ni falso que, si estos sucesos hubiesen ocurrido, yo

[1] Raverat.

nunca habría existido. Aunque siempre puedo preguntar, «¿Habría existido yo?», esto habría sido aquí una pregunta vacía.

Estas últimas afirmaciones son polémicas. Como quiero que mi tesis de dependencia temporal no lo sea, dejaré aparte estos casos. La tesis puede convertirse en

> (TD2) Si una persona particular no hubiese sido concebida en el espacio de un mes alrededor del momento en que de hecho fue concebida, de hecho nunca habría existido.

Afirmo que esto es verdadero *de hecho*. No afirmo que sea *necesariamente* verdadero. Las diferentes concepciones de este tema hacen declaraciones rivales en torno a lo que es necesario. Es porque yo afirmo menos por lo que mi afirmación no es polémica. Los que están en desacuerdo sobre lo que *podría* haber ocurrido pueden estar de acuerdo sobre lo que *de hecho habría* ocurrido. Como mantendré, los que sostienen todas las concepciones que son plausibles estarían de acuerdo conmigo.

Estas concepciones hacen declaraciones sobre las *propiedades necesarias* de cada persona concreta. Algunas de las propiedades necesarias de una persona las tienen todas: se trata de las propiedades que son necesarias para ser una persona. Lo que nos interesa a nosotros aquí son las propiedades necesarias *distintivas* de cada persona particular. Supongamos que afirmo que *P* es una de las propiedades necesarias distintivas de Kant. Esto significa que Kant no podría haber carecido de P, y que sólo Kant podría haber tenido P.

De acuerdo con

> *La Tesis del Origen*, cada persona tiene esta propiedad necesaria distintiva: haberse originado del par concreto de células del que esta persona se originó de hecho [2].

Esta propiedad no puede ser *completamente* distintiva. Todo par de gemelos se originaron *los dos* de un par de células semejante. Y

[2] Véase Kripke, Bogen, Forbes (1).

todo óvulo fecundado podría haberse escindido más adelante, y haber dado origen a gemelos. La tesis del origen tiene que ser revisada para hacer frente a este problema. Pero no necesito discutir esta revisión. Basta para mis propósitos con que, según esta tesis, Kant no podría haberse originado a partir de un par de células diferente. Es irrelevante que, puesto que puede haber gemelos, sea falso que *sólo* Kant podría haberse originado a partir de este par de células.

Los que mantienen la tesis del origen aceptarían mi afirmación de que, si Kant no hubiera sido concebido en el espacio de un mes alrededor del momento en que fue concebido, de hecho nunca habría existido. Si no hubiera sido concebido ese mes, no se habría originado de hecho ningún niño del par de células concreto del que se originó él. (Esta afirmación da por supuestas cosas tanto acerca de las propiedades necesarias distintivas de este par de células, como sobre el sistema reproductivo humano. Pero se trata de asunciones que no son polémicas.)

De acuerdo con otras determinadas concepciones, Kant podría haberse desarrollado a partir de un par de células diferente. Según

> *La Concepción Cartesiana Monótona*, Kant fue un Ego Cartesiano concreto, que *no* tenía propiedades necesarias distintivas.

Según esta concepción, la identidad de una persona no tiene conexiones con sus características físicas y mentales. Kant podría haber sido yo, y viceversa, aunque si hubiera ocurrido esto nadie habría notado ninguna diferencia. Es en el peor de los casos levemente polémico declarar, como hice yo, que deberíamos rechazar *esta* versión de la Concepción Cartesiana.

Hay otras dos concepciones que están estrechamente relacionadas entre sí. Según

> *La Concepción Descriptiva*, cada persona tiene varias propiedades necesarias distintivas. Estas son las propiedades distintivas más importantes de esta persona, y no incluyen haberse originado a partir de un par concreto de células.

En el caso de Kant, estas propiedades incluirían haber sido el autor de ciertos libros. Una versión de esta concepción no afirma

que Kant tenga que haber tenido *todas* estas propiedades. Cualquiera con la mayoría de estas propiedades habría sido Kant.

Según

> *La Concepción del Nombre Descriptivo*, el nombre de toda persona significa «la persona que...». Para nosotros ahora, «Kant» significa «la persona que escribió la *Crítica de la Razón Pura*, etc.». Las propiedades necesarias de una persona particular son aquellas que se enumerarían cuando explicáramos el significado del nombre de esa persona.

Tanto esta como la Concepción Descriptiva podrían combinarse con la otra versión del Cartesianismo. De Kant se podría decir que es el Ego Cartesiano cuyas propiedades necesarias distintivas incluyen la autoría de ciertos libros. Pero ninguna de las dos concepciones descriptivas tiene por qué añadir esta afirmación [3].

Una objeción a las Concepciones Descriptivas es que la vida de cada persona podría haber sido muy diferente. Kant podría haber muerto en su cuna. Como esto es posible, el haber escrito ciertos libros no puede ser una de las propiedades necesarias de Kant.

Hay una respuesta a esta objeción que se retira a una declaración más débil. Se podría decir:

> Aunque esta propiedad no es necesaria, sí que es distintiva. Kant podría no haber escrito estos libros. Pero, en toda posible historia en la que una persona sola escribiera estos libros, esta persona habría sido Kant.

No tengo necesidad de discutir si esta, o cualquier otra respuesta, anula esta objeción. Aunque la objeción pueda anularse, mi Tesis de Dependencia Temporal es verdadera.

Según las dos Concepciones Descriptivas, Kant podría haberse originado a partir de un par de células diferente, o incluso podría haber tenido diferentes padres. Esto habría ocurrido si la madre de

[3] Véase la discusión de estos puntos de vista, y las referencias indicadas, en Kripke.

Kant no hubiera concebido un hijo cuando lo concibió a él, y alguna otra pareja hubiera concebido un hijo que más adelante escribiera la *Crítica de la Razón Pura*, etc. Según las Concepciones Descriptivas, este niño habría sido Kant. No le habrían *llamado* Kant, pero esto no preocupa a los que mantienen estas concepciones. Afirmarían que, si hubiera ocurrido esto, Kant habría tenido diferentes padres y además un nombre diferente.

Aunque piensan que esto podría haber ocurrido, la mayoría de los que mantienen las Concepciones Descriptivas aceptarían mi afirmación de que *no* habría ocurrido *de hecho*. Si afirmaran que *habría* ocurrido, tendrían que aceptar una versión radical de la idea de Tolstoy, plasmada en el epílogo de *Guerra y Paz*, de que la historia no depende de las decisiones que toman las personas concretas. Según esta concepción, si la madre de Napoleón no hubiera tenido hijos, la historia habría proporcionado un «Napoleón sustituto», que habría invadido Rusia en 1812. Y si la madre de Kant no hubiera tenido hijos, la historia habría proporcionado otro autor de la *Crítica de la Razón Pura*. Esta idea es demasiado inverosímil como para que valga la pena discutirla.

Hay otro modo en que los que mantienen las Concepciones Descriptivas podrían rechazar mi tesis. Podrían decir que las propiedades necesarias de Kant eran mucho menos distintivas. Podrían por ejemplo ser simplemente: ser el primer hijo de su madre. Esta afirmación hace frente a la objeción de que la vida de cada persona podría haber sido diferente. Pero es también demasiado inverosímil como para que valga la pena discutirla. Yo soy el segundo de los tres hijos de mi madre. La afirmación referida implica el absurdo de que, si mi madre no hubiera concebido ningún hijo cuando de hecho me concibió a mí, yo habría sido mi hermana más joven.

Consideremos a continuación la historia posible en la que las Concepciones Descriptivas parecen de lo más plausible. Supongamos que la madre de Kant no hubiera concebido un hijo cuando lo concibió a él, y que un mes después concibió un hijo que era exactamente como Kant. Este niño se habría originado a partir de un par de células diferente; pero por una coincidencia asombrosa, de una clase que en la realidad nunca ocurre, este niño habría tenido todos los genes de Kant. Y supongamos que, aparte del hecho y de los efectos de nacer más tarde, este niño hubiera vivido una vida igual que la de Kant, escribiendo la *Crítica de la Razón Pura*, etc.

Según las Concepciones Descriptivas, este niño habría sido Kant. Los que mantienen la Tesis del Origen podrían objetar:

> Kant era una persona concreta. En tu historia posible imaginaria, no has demostrado que te estés refiriendo a *esta* persona concreta. En esta historia imaginaria, habría habido alguien que fue *exactamente como* esta persona. Pero la igualdad exacta no es lo mismo que la identidad numérica, como lo demuestran cualesquiera dos cosas exactamente iguales.

Estas observaciones explican por qué la Tesis del Origen se refiere al par concreto de células a partir del cual se originó una persona.

Una quinta concepción hace también una referencia directa como esta. Según

> *La Concepción de la Variación hacia Atrás*, esta referencia no necesita hacerse al punto de origen, o a las células a partir de las que se originó una persona. Puede hacerse a cualquier momento de la vida de esta persona. Al hacer esta referencia, podemos describir cómo podría haber tenido esta persona un origen diferente [4].

Consideremos a un partidario de esta concepción que, en 1780, asiste a una de las clases de Kant. Esta persona podría afirmar:

> Kant es la persona que está *allí*. Kant podría haber tenido unos padres diferentes y haber vivido una vida diferente hasta el pasado reciente. Para que esto haya sido lo que ocurrió, todo lo que se necesita es que esta vida diferente hubiera conducido a Kant a estar ahora allí.

Esta concepción tiene que hacer algunas afirmaciones suplementarias. Pero hace frente a la objeción de que, para justificar una

[4] Saqué la sugerencia de D. Wiggins, que también cita a H. Ishiguro.

declaración de identidad, necesitamos más que similitud. Los que mantienen la Tesis del Origen necesitan por consiguiente una objeción diferente para plantearle a la Concepción de la Variación hacia Atrás. Para mis propósitos, no tengo que decidir entre estas concepciones.

Según la Concepción de la Variación hacia Atrás, Kant podría haber tenido un origen diferente. Pero los que la mantienen aceptarían mi afirmación de que, de hecho, esto no habría ocurrido. Concederían que, si Kant no hubiese sido concebido en el espacio de un mes alrededor del momento en que fue concebido, de hecho no habría existido nunca.

He descrito ahora todas las opiniones sobre nuestra identidad en las diferentes historias posibles [5]. Discuto cómo están relacionadas estas ideas con las diferentes concepciones de la identidad a través del tiempo [6]. Según todas las concepciones

[5] Hay personas que afirman que no hay propiedades esenciales distintivas. Esto implica que, para cada una de estas personas, es una cuestión vacía la de si, en caso de que sus padres nunca se hubiesen casado, *ella* habría existido nunca. Aunque pienso que podría haber cuestiones vacías acerca de nuestra identidad, dudo que, tras reflexionar, estas personas siguieran creyendo que *esta* cuestión es vacía. (En el caso especial de los gemelos univitelinos, hay algunas cuestiones vacías aquí. Espero discutir este asunto en otra parte.)

[6] ¿Cuál es la relación entre este tema y el de la identidad personal a través del tiempo? Según la Tesis del Origen, es una propiedad esencial de cada persona haberse originado a partir de un huevo fertilizado particular. Este modo de ver las cosas podría combinarse con el Criterio Físico. Según la versión más convincente de este criterio, una propiedad esencial de cada persona es que tiene lo bastante de su cerebro particular como para dar soporte a una vida consciente plena. Podría afirmarse que es una propiedad esencial de cualquier cerebro particular haberse originado a partir de un huevo fertilizado particular.

El Criterio Físico no necesita combinarse con la Tesis del Origen. Alguien que crea en este criterio podría aceptar la Concepción de la Variación hacia Atrás. Describí cómo, según ella, Tolstoy podría haber tenido un origen diferente. En esta posible y diferente historia imaginaria, se habría cumplido el Criterio Físico. De manera similar, podríamos aceptar la Tesis del Origen pero rechazar el Criterio Físico. Podríamos combinar la Tesis del Origen con las versiones amplias del Criterio Psicológico. Entonces pensaríamos que es esencial para mí haberme ori-

plausibles, mi Tesis de la Dependencia Temporal es verdadera. Esta tesis se aplica a todos. Tú fuiste concebido en un determinado momento. Es de hecho verdadero que, si no hubieras sido concebido en el espacio de un mes alrededor de ese momento, *tú* no habrías existido nunca.

120. LAS TRES CLASES DE ELECCIÓN

A no ser que nosotros, o un desastre global, destruyamos al género humano, habrá personas que ahora no existen viviendo más adelan-

ginado de un par particular de células, y, por tanto, haber comenzado a vivir en un cuerpo particular. Pero también pensaríamos que mi vida podría continuar en un cuerpo diferente. Esto ocurriría, por ejemplo, si fuese teletransportado.

Considérese a continuación el Criterio Psicológico, que apela a la continuidad psicológica. Los que creen en él podrían estar de acuerdo en que mi vida podría haber ido de forma muy diferente. Y entre yo en mi vida real y yo mismo en esta diferente vida posible podría haber muy poca continuidad psicológica. Supongamos que mis padres me hubieran llevado a Italia cuando yo tenía tres años, y que me hubiera convertido en italiano. Entonces habría la siguiente relación entre yo ahora y yo mismo como podría haber sido ahora. Tanto en mi vida real como en esta diferente vida posible yo sería ahora psicológicamente continuo conmigo mismo en mi vida real cuando tenía tres años. Esta relación sería débil. Hay pocas conexiones psicológicas directas entre yo ahora y yo mismo a los tres años. Si comparamos mi vida real con esta diferente vida posible, podría ser que las dos vidas no contuviesen en la edad adulta ni siquiera un recuerdo común.

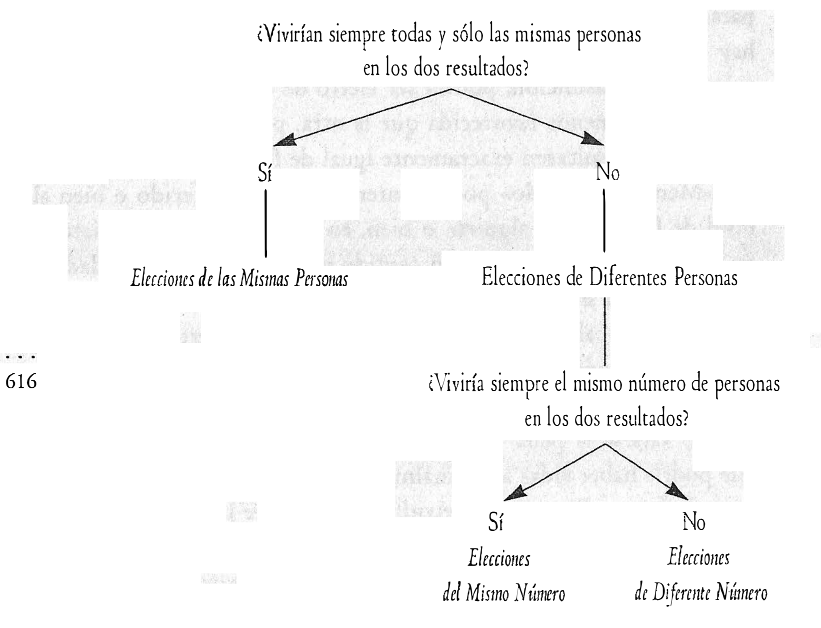
Algunos de los que creen en el Criterio Psicológico podrían afirmar que estas no son dos vidas posibles de la misma persona particular —que el niño que habría marchado a Italia a la edad de tres años no habría sido yo sino una persona diferente—. Esta afirmación revisa nuestra concepción corriente de la identidad personal.

Tal revisión no sólo es técnicamente difícil, tampoco es necesaria. Se puede plantear la cuestión de una manera más simple. Si me hubiese ido a Italia a los tres años, mi vida habría sido muy diferente. Y podemos creer que este hecho tiene varias implicaciones prácticas y morales. Pero esta creencia no necesita formularse negando la identidad entre yo en mi vida real y yo mismo en esta diferente vida posible. Podemos admitir que esta relación es la de la identidad. Podemos plantear nuestra cuestión afirmando que, cuando estamos comparando tales vidas posibles muy diferentes, el hecho de la identidad personal no tiene su significación corriente. Véase Adams (3).

te. Son las *personas futuras*. La ciencia le ha dado a nuestra generación una gran capacidad tanto para afectar a esas personas como para predecir estos efectos.

Dos clases de efectos hacen surgir cuestiones enigmáticas. Nosotros podemos influir en las identidades de las personas futuras, o en *quién* va a ser la gente que vivirá más adelante. Y podemos influir en el número de personas futuras. Estos efectos nos dan diferentes clases de elección.

Al comparar dos actos cualesquiera, podemos preguntar:



Las Elecciones de Diferente Número afectan tanto al número cuanto a las identidades de las personas futuras. Las Elecciones del Mismo Número afectan a las identidades de las personas futuras, pero no a su número. Las Elecciones de las Mismas Personas no afectan a ninguno de los dos.

121. ¿QUÉ PESO DEBERÍAMOS DAR A LOS INTERESES DE LAS PERSONAS FUTURAS?

La mayor parte de nuestro pensamiento moral versa sobre Elecciones de las Mismas Personas. Como sostendré, estas elecciones no son tan numerosas como la mayoría de nosotros asume. Muchísimas elecciones nuestras tendrán de hecho algún efecto tanto sobre las identidades cuanto sobre el número de las personas futuras. Pero en la mayoría de estos casos, como no podemos predecir cuáles serían los efectos concretos, estos efectos pueden ser moralmente ignorados. Podemos tratar estos casos como si fueran Elecciones de las Mismas Personas.

En algunos casos podemos predecir que un acto irá o podrá ir en contra de los intereses de las personas futuras. Esto puede ser cierto cuando hagamos una de las Mismas Personas. En tal caso, elijamos lo que elijamos, todas y solamente las mismas personas vivirán en todo momento. Algunas de estas personas serán personas futuras. Como ellas existirán elijamos lo que elijamos, podemos perjudicarlas o beneficiarlas muy claramente.

Supongamos que dejo un cristal roto entre la maleza de un bosque. Cien años después el cristal hiere a un niño. Mi acto daña a ese niño. Si yo hubiera enterrado con precaución el cristal, el niño habría caminado por el bosque sin sufrir daño alguno.

¿Supone una diferencia moral el que el niño a quien hago daño no exista ahora?

Para una concepción, los principios morales incluyen sólo a las personas que puedan *corresponder*, o dañarse y beneficiarse mutuamente. Si a mí este niño no me puede ni dañar ni beneficiar, como podemos suponer plausiblemente, el daño que le causo no tiene importancia moral. Doy por sentado que deberíamos rechazar esta concepción [7].

[7] Por las razones dadas por el trabajo de Brian Barry, «Circumstances of Justice and Future Generations» [«Las circunstancias de la justicia y las generaciones futuras»], en Sikora y Barry.

Algunos autores afirman que, mientras que debemos preocuparnos por los efectos en las personas futuras, estamos moralmente justificados si nos preocuparnos menos por los efectos en el futuro más distante. Esta es una concepción común en economía del bienestar y en los análisis coste-beneficio. Según ella, podemos *descontar* los efectos más remotos de nuestros actos y de nuestras políticas, a razón de *n* por ciento por año. Esto se llama la *Tasa de Descuento Social*.

Supongamos que estamos considerando cómo deshacernos de un modo seguro de la materia radioactiva llamada *desechos nucleares*. Si creemos en la Tasa de Descuento Social, nos preocuparemos de la seguridad sólo en el futuro más próximo. No nos afectará el hecho de que haya desechos nucleares que permanecerán radioactivos durante miles de años. A una tasa de descuento del cinco por ciento, una muerte el año próximo cuenta como más de mil millones de muertes dentro de 500 años. Según este modo de pensar, las catástrofes en el futuro más lejano pueden considerarse ahora moralmente insignificantes.

Como sugiere este caso, la Tasa de Descuento Social no es defendible. Lo remoto en el tiempo correlaciona más o menos con algunos hechos importantes, como la previsibilidad. Pero, como sostengo en el Apéndice F, estas correlaciones son demasiado toscas como para justificar la Tasa de Descuento Social. La importancia moral presente de los sucesos futuros *no* declina a razón de *n* por ciento por año. Lo remoto en el tiempo no tiene, en sí mismo, más significación que lo remoto en el espacio. Supongamos que disparo una flecha a un bosque lejano, donde hiere a una persona. Si yo hubiera sabido que podría haber alguien en ese bosque, sería culpable de flagrante negligencia. Como la persona está lejos, no puedo identificar a quién daño. Pero esto no es excusa. Ni tampoco lo es que la persona esté lejos. Deberíamos decir lo mismo sobre los efectos en personas que son temporalmente remotas.

122. EL HIJO DE UNA JOVEN

Las personas futuras no son, por lo menos en un aspecto, como las personas lejanas. Podemos influir en su identidad. Y muchos de nuestros actos tienen este efecto.

Este hecho da lugar a un problema. Antes de describirlo, repetiré algunas observaciones preliminares. Asumo que una persona puede resultar menos favorecida que otra, de formas moralmente significativas, y en más o en menos. Pero no asumo que estas comparaciones puedan ser precisas, ni siquiera en principio. Asumo que hay sólo una posibilidad de comparación parcial o aproximada. A tenor de esta asunción, podría ser cierto de dos personas que ninguna resulte menos favorecida que la otra, pero esto no implicaría que las dos resultasen exactamente igual de favorecidas.

«Menos favorecido» podría entenderse como referido o bien al nivel de felicidad de alguien, o bien, en un sentido más estricto, a su nivel de vida, o bien, en un sentido más amplio, a su calidad de vida. Como es el sentido más amplio de todos, a menudo usaré la expresión «la calidad de vida». También diré de ciertas vidas «que vale la pena vivirlas», o que son «dignas de ser vividas». Esta descripción la pueden ignorar los que creen que no podría haber vidas que no valiese la pena vivirlas. Pero yo creo, como muchos otros, que podría haber vidas así. Finalmente, voy a ampliar el uso corriente de la frase «digna de ser vivida», o «que vale la pena vivirla». Si una de dos personas tuviera una calidad de vida más baja, diré que su vida es, en esta medida, «menos digna de ser vivida».

Cuando consideramos a personas futuras, tenemos que responder a dos preguntas:

> (1) Si hacemos que alguien exista, y ese alguien va a tener una vida digna de ser vivida, ¿con ello beneficiamos a esa persona?
> (2) ¿Beneficiamos también a esa persona si algún acto nuestro es una parte, remota pero necesaria, de la causa de su existencia?

Estas son preguntas difíciles. Si contestamos Sí a las dos, diré que creemos que *causar que se exista puede beneficiar*.

Algunos contestarían Sí a (1) pero No a (2). Estas personas dan su segunda respuesta porque usan «beneficio» en su sentido corriente. Como sostuve en la Sección 25, debemos ampliar, para propósitos morales, nuestro uso de «beneficio». Si contestamos Sí a (1) deberíamos contestar Sí a (2).

Muchos responden No a estas dos preguntas. Estas personas podrían decir: «Beneficiamos a alguien si resulta que, si no hubiéramos hecho lo que hicimos, habría sido peor para esta persona. Si no hubiéramos hecho existir a alguien, esto *no* habría sido peor para esta persona».

Yo pienso que, mientras que es defendible responder No a estas dos preguntas, también lo es responder Sí a las dos. Para los que dudan de esta segunda creencia he escrito el Apéndice G. Como pienso que es defendible tanto afirmar como negar que hacer existir pueda beneficiar, discutiré las implicaciones de ambas opiniones.

Consideremos

> *La Muchacha de 14 Años.* Esta muchacha decide tener un hijo. Como es tan joven, le da a su hijo un mal comienzo en la vida. Aunque esto tendrá efectos negativos durante la vida de su hijo, la vida de éste será, previsiblemente, digna de ser vivida. Si esta muchacha esperara varios años, habría tenido un hijo diferente, a quien habría dado un mejor comienzo en la vida.

Como tales casos se están haciendo comunes, plantean un problema práctico [8]. Y también un problema teórico.

Supongamos que intentamos convencer a esta muchacha de que debe esperar. Le dijimos: «Si tienes un hijo ahora, pronto lo lamentarás. Si esperas, será mejor para ti». Ella contestó: «Es asunto mío. Aunque haga lo que es peor para mí, tengo derecho a hacer lo que quiero».

[8] Véase *Teenage Pregnancy in a Family Context: Implications for Policy Decisions* [*El embarazo de adolescentes en un contexto familiar: implicaciones para decisiones políticas*], editado por Teodora Ooms (de soltera Parfit), Temple University Press, Philadelphia, 1981.

Nosotros replicamos: «No sólo es asunto tuyo. No deberías pensar sólo en ti misma, sino también en tu hijo. Será peor para él si lo tienes ahora. Si lo tienes más adelante, le darás un mejor comienzo en la vida».

No conseguimos convencerla. Tuvo el niño a los 14 años, y, como predijimos, le dio un mal comienzo en la vida. ¿Estábamos en lo correcto cuando afirmábamos que su decisión fue peor para su hijo? Si ella se hubiese esperado, este niño concreto nunca hubiera existido. Y, no obstante su mal comienzo, su vida es digna de vivirse. Supongamos en primer lugar que *no* creemos que hacer que se exista pueda beneficiar. Deberíamos preguntar: «si alguien vive una vida digna de ser vivida, ¿esto es para esta persona peor que si nunca hubiera existido?». Nuestra respuesta tiene que ser No. Supongamos después que pensamos que hacer que se exista *puede* beneficiar. Según este modo de ver, la decisión de esta muchacha beneficia a su hijo.

Según ambas opiniones, la decisión de la muchacha no fue peor para su hijo. Cuando nos damos cuenta de esto, ¿cambiamos de opinión sobre esta decisión? ¿Dejamos de pensar que habría sido mejor que la muchacha se hubiera esperado, para poder dar a su primer hijo un mejor comienzo en la vida? Sigo teniendo esta creencia, como la mayor parte de las personas que consideran este caso. Pero no podemos defenderla del modo natural que sugerí. No podemos decir que la decisión de la muchacha fue peor para su hijo. ¿Cuál es la objeción a su decisión? Surge esta pregunta porque, en los diferentes resultados, nacerían diferentes personas. Por eso llamaré a este problema el *Problema de la No-Identidad* [9].

Puede decirse:

> En cierto sentido, la decisión de la muchacha *fue* peor para su hijo. Al tratar de convencerla de que no tenga un hijo ahora, podemos usar la frase «su hijo» y el pronombre «él» para incluir a *cualquier* hijo que ella pudiera tener. Estas palabras no tienen por qué refe-

[9] Kavka ha denominado a este problema *la paradoja de los individuos futuros.* Véase Kavka (4).

rirse a un niño en particular. Podemos afirmar realmente: «Si la muchacha no tiene su hijo ahora, sino que se espera y lo tiene más adelante, *él* será un niño diferente». Usando estas palabras así, podemos explicar por qué sería mejor que la muchacha se espere. Podemos decir:

(A) La objeción a la decisión de la muchacha es que probablemente será peor para su hijo. Si se esperara, probablemente le daría un mejor comienzo en la vida.

Aunque realmente podamos hacer esta afirmación, *no* explica la objeción a la decisión de la muchacha. Cosa que queda clara después de que ella ha tenido su hijo. La expresión «su hijo» se refiere ahora naturalmente a este niño en particular. Y la decisión de la muchacha *no* fue peor para *este* niño. Aunque haya un sentido en que (A) es verdadero, (A) no apela a un principio moral corriente.

Según uno de nuestros principios corrientes, es una objeción a la elección de una persona el que vaya a ser peor para, o vaya a ir en contra de los intereses de, cualquier otra persona concreta. Si afirmamos que la decisión de la muchacha fue peor para su hijo, no podemos estar diciendo que fue peor para una persona en particular. No podemos afirmar, del hijo de la muchacha, que la decisión de ella fue peor para *él*. Tenemos que admitir que, en la afirmación (A), las palabras «su hijo» no refieren a su hijo. (A) no trata sobre lo que es bueno o malo para cualquiera de las personas concretas que alguna vez vivan. (A) apela a un principio nuevo, que tiene que explicarse y justificarse.

Si (A) parece apelar a un principio corriente, esto es porque tiene dos sentidos. Aquí va otro ejemplo. Un general hace gala de maestría militar si, en muchas batallas, siempre hace suyo el bando ganador. Pero hay dos modos de hacer esto. Él podría ganar victorias. O él podría cambiar de bando siempre que está a punto de perder. Un general no hace gala de maestría militar si es sólo en el segundo sentido que siempre hace suyo el bando ganador.

¿A qué principio apela (A)? Deberíamos formularlo de manera que mostrara la clase de elección a la que se aplica. Son las Elecciones del Mismo Número, que afectan a las identidades de las personas futuras, pero no a su número. Podríamos sugerir

> *La Tesis de la Calidad del Mismo Número*, o C: Si en cada uno de dos resultados posibles vivirían siempre el mismo número de personas, sería peor si los que viven resultan menos favorecidos, o tienen una calidad de vida más baja, que los que hubieran vivido.

Esta tesis es plausible. E implica lo que pensamos sobre la muchacha de 14 años. El hijo que tiene ahora probablemente resultará menos favorecido que el hijo que podría haber tenido más adelante, puesto que este otro niño habría tenido un mejor comienzo en la vida. Si esto es verdadero, C implica que este es el peor de los dos resultados. C implica que habría sido mejor si esta muchacha se hubiera esperado y hubiera tenido un hijo más adelante.

Puede que no nos atrevamos a decir, del hijo real de la muchacha, que habría sido mejor que nunca hubiera existido. Pero, si antes afirmamos que sería mejor que la muchacha se esperara, esto es lo que tenemos que decir. No podemos de una manera consistente hacer una afirmación y negar la misma afirmación después. Si (1) en 1990 *sería* mejor que la muchacha se esperara y tuviera el niño más tarde, entonces (2) en 2020 *habría sido mejor* que se hubiera esperado y hubiera tenido el niño más tarde. Y (2) implica (3) que habría sido mejor que el niño que existía no hubiera sido su hijo real. Si no podemos aceptar (3) tenemos que rechazar (1).

Después de reflexionar sugiero que podemos aceptar (3). Creo que, si *yo* fuera el hijo real de esta muchacha, podría aceptar (3). (3) no implica que mi existencia sea *mala*, o intrínseca y moralmente indeseable. La afirmación es meramente que, como un niño nacido más tarde probablemente habría tenido una vida mejor que la mía, habría sido mejor que mi madre se hubiera esperado, y hubiera tenido un niño más tarde. Esta afirmación no tiene por qué implicar que yo deba racionalmente lamentar que mi madre *me* tuviera, o que ella deba racionalmente lamentarlo. Como habría sido mejor que se hubiera esperado, quizás deba ella tener algún remor-

dimiento de orden moral. Y es probablemente verdadero que produjo el peor resultado para sí misma. Pero, aunque esto sea así, no demuestra que ella deba racionalmente lamentar su acto, una vez consideradas todas las cosas. Si ella me quiere a mí, su hijo real, esto basta para bloquear la afirmación de que ella es irracional si no tiene tal remordimiento [10]. Aunque implique una afirmación como (3), concluyo que podemos aceptar C.

Aunque C sea plausible, no resuelve el Problema de la No-Identidad. C incluye sólo los casos en que, en los diferentes resultados, el mismo número de personas viviría en todo momento. Necesitamos una tesis que incluya casos en que, en los diferentes resultados, vivirían en todo momento diferentes números. El Problema de la No-Identidad puede surgir en estos casos.

Puesto que C es limitada, podría justificarse de diversos modos diferentes. Hay varios principios que implican C, pero entran en conflicto cuando los aplicamos a Elecciones de Diferente Número. Tendremos que decidir cuál de estos principios, o cuál conjunto de principios, debemos aceptar. Llamemos a lo que debemos aceptar *Teoría X*. X resolverá el Problema de la No-Identidad en las Elecciones de Diferente Número. Y X nos dirá cómo debería justificarse C, o cómo debería explicarse más completamente.

En el caso de la muchacha de 14 años, no estamos forzados a apelar a C. Hay otros hechos a los que podríamos apelar, como los efectos sobre otras personas. Pero el problema puede surgir en una forma más pura.

123. Cómo la disminución de la calidad de vida podría no ser peor para nadie

Supongamos que elegimos entre dos políticas sociales o económicas. Y supongamos que, en una de las dos políticas, el nivel de vida sería ligeramente más alto en el próximo siglo. Este efecto

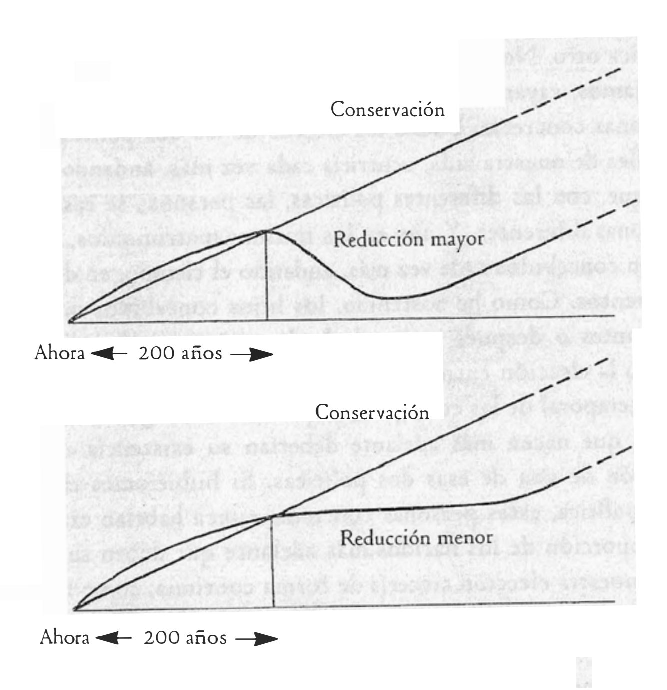
[10] Sigo a Adams (3).

implica otro. No es cierto que, cualquiera que sea la política que escojamos, vayan a existir en el futuro más distante las mismas personas concretas. Dados los efectos de las dos políticas en los detalles de nuestra vida, ocurriría cada vez más, andando el tiempo, que, con las diferentes políticas, las personas se casaran con personas diferentes. Y, aun en los mismos matrimonios, los hijos serían concebidos cada vez más, andando el tiempo, en diferentes momentos. Como he sostenido, los hijos concebidos más de un mes antes o después serían de hecho niños y niñas diferentes. Como la elección entre nuestras dos políticas afectaría a la posición temporal de las concepciones posteriores, algunas de las personas que nacen más adelante deberían su existencia a nuestra elección de una de esas dos políticas. Si hubiéramos elegido la otra política, estas personas concretas nunca habrían existido. Y la proporción de los nacidos más adelante que deben su existencia a nuestra elección crecería de forma continua, como las ondas en un estanque. Podemos asumir de manera plausible que, uno o dos siglos después, no habría nadie viviendo en nuestra comunidad que hubiera nacido cualquiera que hubiera sido la política elegida. (Puede ser útil pensar en esta pregunta: ¿cuántos de nosotros podrían afirmar verdaderamente, «Aunque los trenes y los coches no se hubieran inventado nunca, yo sin embargo habría nacido»?)

¿Cómo genera esto un problema? Consideremos

> *La Reducción.* Como comunidad, tenemos que elegir entre reducir o conservar ciertas clases de recursos. Si elegimos la Reducción, la calidad de vida durante los próximos dos siglos sería ligeramente más alta de lo que habría sido si hubiéramos elegido la Conservación. Pero más tarde sería, durante muchos siglos, mucho más baja de lo que lo habría sido si hubiéramos elegido la Conservación. La razón sería que, al comienzo de este período, la gente tendría que encontrar alternativas para los recursos que nosotros habíamos reducido. Vale la pena distinguir dos versiones de este caso. Los efectos de las diferentes políticas serían como los que se muestran en la siguiente página





Nunca podríamos saber, con tanto detalle, que estos serían los efectos de las dos políticas. Pero esto no es objeción ninguna a este caso. Efectos similares serían previsibles algunas veces. Tampoco importa que este caso imaginario sea artificialmente simple, puesto que esto lo único que hace es aclarar las cuestiones relevantes.

626

Supongamos que elegimos la Reducción, y que esto tiene uno de los dos efectos mostrados en mi diagrama. ¿Es nuestra elección peor para alguien?

Como elegimos la Reducción, millones de personas tienen, durante varios siglos, una calidad de vida muy inferior. Esta calidad de vida es muy inferior no en comparación con la de ahora sino en comparación con la que habrían tenido si hubiéramos elegido la conservación. Las vidas de estas personas son dignas de ser vividas; y, si hubiéramos elegido la Conservación, estas personas concretas nunca habrían existido. Supongamos que no damos por sentado que causar que se exista pueda beneficiar. Deberíamos preguntar: «si las personas concretas viven vidas que vale la pena vivir, ¿esto es para ellas peor que si nunca hubieran existido?». Nuestra respuesta tiene que ser No. Supongamos a continuación que damos por sentado

que causar que se exista puede beneficiar. Como las vidas de estas personas futuras serán dignas de vivirse, y ellas nunca habrían existido si hubiéramos elegido la Conservación, nuestra elección de la Reducción no es que no sólo no sea peor para ellas: es que les *beneficia.*

A tenor de ambas respuestas, nuestra elección no será peor para estas personas futuras. Además, cuando entendemos el caso, sabemos que esto es verdadero. Sabemos que, aunque baje enormemente la calidad de vida por varios siglos, nuestra elección no será peor para nadie que alguna vez viva.

¿Representa esto una diferencia moral? Hay tres opiniones. Podría representar toda la diferencia, o alguna diferencia, o ninguna diferencia. Podría no haber objeción ninguna a nuestra elección, o alguna objeción, o la objeción ser igual de fuerte.

Hay quienes creen que *lo que es malo tiene que ser malo para alguien.* Según esta opinión, no hay objeción que plantear a nuestra elección. Como no será mala para nadie, nuestra elección no puede tener un mal efecto. La gran disminución de la calidad de vida no proporciona ninguna razón moral para no elegir la Reducción.

Algunos autores aceptan esta conclusión [11]. Pero es muy poco plausible. Antes de considerar casos de esta clase, podemos 627
aceptar la idea de que lo que es malo tiene que ser malo para alguien. Pero el caso de la Reducción demuestra, creo, que tenemos que rechazar esta idea. La gran disminución de la calidad de vida tiene que proporcionar *alguna* razón moral para no elegir la Reducción. Esto lo piensa la mayoría de los que consideran casos de esta clase.

Si esto es lo que pensamos, deberíamos hacer dos preguntas:

(1) ¿Cuál es la razón moral para no elegir la Reducción?
(2) ¿Representa una diferencia moral el que esta disminución de la calidad de vida no sea peor para nadie? ¿Sería *peor* este efecto, teniendo así un peso moral mayor, si *fuera* peor para personas concretas?

[11] Véase T. Schwartz, «Obligations to Posterity» [«Obligaciones con la posteridad»], en Sikora y Barry.

A nuestra necesidad de contestar a (1), y a otras preguntas parecidas, la llamo el Problema de la No-Identidad. Este problema surge porque las identidades de las personas en el futuro lejano pueden ser afectadas muy fácilmente. Algunos piensan que este problema es una objeción de poca monta. Esta reacción no está justificada. El problema surge a consecuencia de ciertos hechos superficiales sobre nuestro sistema reproductivo. Pero, aunque surja de un modo superficial, es un problema real. Cuando elegimos entre dos políticas sociales o económicas como las que he descrito, *no* es *verdad* que, en el futuro lejano, vayan a existir las mismas personas, elijamos lo que elijamos. Por eso *no* es *verdad* que una elección como la de la Reducción vaya en contra de los intereses de las personas futuras. No podemos despachar este problema fingiendo que esto *es* verdadero.

Respondemos parcialmente a la pregunta (1) si apelamos a C. Según esta afirmación, si los números fuesen los mismos, sería peor si los que viven tuvieran una calidad de vida más baja que los que habrían vivido. Pero el problema puede surgir en casos en que, en los diferentes resultados, habría diferentes números de personas. Para incluir estos casos tenemos necesidad de la Teoría X. Sólo X explicará cómo debería justificarse C, y sólo X proporcionará una solución completa a nuestro problema.

124. Por qué una apelación a los derechos no puede resolver del todo el problema

¿Podemos resolver nuestro problema apelando a los derechos de las personas? Reconsideremos a la muchacha de 14 años. Teniendo a su hijo tan joven le da un mal comienzo en la vida. Podría afirmarse: «La objeción a la decisión de la muchacha es que ella vulnera el derecho de su hijo a un buen comienzo en la vida».

Aunque este niño tenga este derecho, podría no haberse realizado. La muchacha podría no haber tenido *este* niño cuando fuese una mujer adulta. Algunos dirían que, como el derecho de este niño no podía ser realizado, no puede afirmarse que la muchacha vulnere este derecho. El objetor podría responder: «Es incorrecto causar que alguien exista si sabemos que esta persona tendrá un derecho que no puede realizarse». ¿Puede ser esta la objeción a la decisión de la muchacha? [13].

Hace unos años, un político británico manifestó su alegría por el hecho de que, el año anterior, se hubiese dado una disminución de los embarazos en las adolescentes. Un hombre de mediana edad escribió enfadado a *The Times*. Había nacido cuando su madre tenía sólo 14 años. Admitía que, como consecuencia de la excesiva juventud de la madre, sus primeros años habían sido duros para ambos. Pero ahora su vida era muy digna de vivirse. ¿Estaba sugiriendo el político que habría sido mejor que él nunca hubiera nacido? Esta sugerencia le parecía monstruosa.

El político estaba sugiriendo esto implícitamente. Según su opinión, habría sido mejor que la madre de este hombre se hubiera esperado varios años antes de tener hijos. Creo que deberíamos aceptar esta opinión. Pero ¿la podemos explicar convincentemente diciendo que el hombre enfadado tenía un derecho que no fue realizado?

Pienso que no. Supongamos que tengo derecho a la intimidad. Te pido que te cases conmigo. Si aceptas, no estás obrando mal al vulnerar mi derecho a la intimidad. Como estoy contento de que obres como lo haces, en lo que a ti respecta yo *renuncio* a este derecho. Una afirmación parecida se aplica al que escribió la indignada carta a *The Times*. Según la sugerencia hecha arriba, este hombre tenía derecho a nacer de una mujer adulta que le diese un buen comienzo en la vida. Su madre actuó incorrectamente porque le hizo existir con un derecho que no se podía realizar. Pero la carta de este hombre demuestra que estaba contento de estar vivo. Niega que su madre obrara mal a causa de lo que le hizo a él. Si hubiéramos afirmado que el acto de ella era incorrecto porque él tiene un derecho que no puede ser realizado, él podría haber dicho, «renuncio a ese derecho». Esto habría socavado nuestra objeción al acto de su madre.

[13] Esta forma de la objeción se sugiere en Tooley.

Habría sido mejor que la madre de este hombre se hubiese esperado. Pero no a causa de lo que ella le hizo a su hijo real, sino a causa de lo que podría haber hecho por cualquier niño o niña que pudiera haber tenido de adulta. La objeción tiene que ser que, si ella se hubiera esperado, podría haber dado a algún otro niño o niña un mejor comienzo en la vida.

Volvamos ahora al Caso de la Reducción. Supongamos que elegimos la Reducción Mayor. Más de dos siglos después, la calidad de vida es mucho más baja de lo que lo habría sido si hubiéramos elegido la Conservación. Pero las personas que entonces vivan tendrán una calidad de vida aproximadamente tan alta como lo será por término medio la nuestra en el próximo siglo. ¿Tienen estas personas derechos a los que pueda apelar un objetor?

Podría decirse que tienen el derecho a su parte de los recursos que hemos reducido. Pero la gente no tiene derechos a una parte de un recurso concreto. Supongamos que reducimos un recurso pero inventamos la tecnología que permitirá a nuestros sucesores, aunque carezcan de este recurso, tener la misma gama de oportunidades. No habría objeción que poner a lo que hemos hecho. Se podría decir, como mucho, que las personas de cada generación tienen el derecho a una gama igual de oportunidades, o a una calidad de vida igualmente alta [14].

Si elegimos la Reducción Mayor, los que vivan más de dos siglos después tendrán menos oportunidades, y una calidad de vida más baja, que algunas generaciones anteriores, y algunas posteriores. Si las personas tienen derecho a oportunidades iguales, y a una calidad de vida igualmente alta, una apelación a estos derechos puede proporcionar una objeción a nuestra elección. Los que vivan más de dos siglos después no podrían bajo ningún concepto haber tenido mayores oportunidades, o una calidad de vida más elevada. Si hubiéramos elegido de otro modo, nunca habrían existido. Como sus derechos no pueden realizarse, tampoco los podemos vulnerar. Pero, como antes, puede objetarse que hacemos que existan personas con derechos que no pueden realizarse.

[14] Véase B. Barry, «Intergenerational Justice in Energy Policy» [«Justicia intergeneracional en política energética»], en Maclean y Brown; véase también Barry (2).

No está claro que esta sea una buena objeción. Si estas personas conociesen los hechos, no lamentarían que actuáramos como actuamos. Si estuvieran contentos de estar vivos, podrían reaccionar como el que escribió a *The Times*. Podrían renunciar a sus derechos. Pero, como no podemos dar por hecho que así es como todos ellos reaccionarían, una apelación a sus derechos puede aportar una objeción a nuestra elección.

¿Esta apelación puede aportar también una objeción a nuestra elección de la Reducción *Menor*? En este caso, los que vivan más de dos siglos después tienen una calidad de vida mucho más alta de la que tenemos ahora nosotros. ¿Podemos afirmar que estas personas tienen *derecho* a una calidad de vida *aún mayor*? Creo que, para toda teoría de derechos plausible, la respuesta sería No.

Será útil eliminar con la imaginación el Problema de la No-Identidad. Supongamos que nuestro sistema reproductivo fuese muy diferente. Supongamos que, cualquiera que fuese la política que siguiéramos, vivirían las mismas personas más de dos siglos después. La objeción a nuestra elección sería entonces que, para conseguir pequeños beneficios para nosotros y nuestros hijos, evitamos que muchas personas futuras reciban beneficios mucho mayores. Como estas personas futuras resultarían más favorecidas que nosotros, no estaríamos actuando injustamente. La objeción a nuestra elección tendría que apelar al Principio de Utilidad.

¿Podría apelar a los derechos esta objeción? Sólo si, como Godwin, presentamos el Utilitarismo como una teoría de los derechos. Según la opinión de Godwin, todo el mundo tiene derecho a recibir lo que el Principio de Utilidad implica que se le debería dar. La mayoría de los que creen en los derechos rechazaría esta opinión. Muchos explican los derechos como lo que *constriñe*, o *limita*, el Principio de Utilidad. Estas personas afirman que es incorrecto vulnerar ciertos derechos, aunque esto incrementara mucho la suma neta de beneficios menos cargas. Según semejante teoría, se le da algún peso al Principio de Utilidad. Como la teoría no es utilitarista, a este principio mejor lo llamamos el ***Principio de Beneficencia***.

Este principio es una parte de esa teoría, y la afirmación de que tenemos ciertos derechos es una parte diferente de la misma. Asumiré que, si creemos en los derechos, este es el tipo de teoría moral que aceptamos.

Volvamos al caso en que eliminamos con la imaginación el Problema de la No-Identidad. Si rechazamos la idea de Godwin, no podríamos poner objeciones a la elección de la Reducción Menor apelando a los derechos de los que vivirán en el futuro lejano. Nuestra objeción apelaría al Principio de Beneficencia. La objeción sería que, para obtener pequeños beneficios para nosotros mismos y nuestros hijos, les negamos, a personas más favorecidas que nosotros, beneficios muchísimo mayores. Al llamar a esto una objeción, no tengo necesidad de afirmar que demuestre que nuestra elección es incorrecta. Estoy afirmando simplemente que, como les negamos a estas personas beneficios muchísimo mayores, esto proporciona *una cierta* razón moral para no hacer esta elección.

Si ahora restablecemos nuestro sistema reproductivo real, esta razón desaparece. Consideremos a las personas que vivirán más de dos siglos más tarde. Nuestra elección de la Reducción Menor no les niega *ningún* beneficio. Si hubiéramos elegido la Conservación, esto no les habría beneficiado, puesto que nunca habrían existido.

Cuando suponemos que no se da el Problema de la No-Identidad, nuestra razón para no hacer esta elección se explica por una apelación, no a los derechos de las personas, sino al Principio de Beneficencia. Cuando restablecemos el Problema de la No-Identidad, esta razón desaparece. Como esta razón apelaba al Principio de Beneficencia, lo que el problema muestra es que este principio es inadecuado, y tiene que revisarse. Necesitamos una mejor explicación de la beneficencia, o lo que llamo Teoría X.

Una parte de nuestra teoría moral apela a la beneficencia, otra parte a los derechos de las personas. Por consiguiente no deberíamos esperar que una apelación a los derechos pudiera llenar el hueco en nuestro inadecuado Principio de Beneficencia. Lo que deberíamos esperar es que, como he afirmado, apelar a los derechos no pueda resolver del todo el Problema de la No-Identidad [15].

125. ¿El hecho de la No-Identidad representa una diferencia moral?

Al tratar de revisar nuestro Principio de Beneficencia —al tratar de encontrar la Teoría X— tenemos que considerar casos en que, en los diferentes resultados, existirían diferentes números de personas. Antes de que nos volvamos a estos casos, podemos preguntar qué pensamos de la otra pregunta que mencioné. Nuestra elección de la Reducción no será peor para nadie. ¿Representa esto una diferencia moral?

Tal vez podamos recordar una época en la que estábamos preocupados por los efectos sobre las generaciones futuras, pero habíamos pasado por alto el Problema de la No-Identidad. Tal vez hayamos pensado que una política como la de la Reducción iría en contra de los intereses de las personas futuras. Cuando vimos que esto era falso, ¿nos volvimos menos preocupados por los efectos sobre las generaciones futuras?

Cuando vi el problema, no me volví menos preocupado. Y lo mismo es cierto de muchas otras personas. Diré que aceptamos la *Tesis de la No-Diferencia.*

Vale la pena considerar un ejemplo diferente:

> *Los Programas Médicos.* Hay dos extrañas enfermedades, *J* y *K*, que no pueden detectarse sin análisis especiales. Si una mujer embarazada contrae la enfermedad J, esto ocasionará que su hijo tenga una cierta discapacidad. Un sencillo tratamiento evitaría este efecto. Si una mujer tiene la enfermedad K cuando concibe a su hijo, esto causará que el niño tenga la misma discapacidad especial. La enfermedad K no puede tratarse, pero siempre desaparece en dos meses.

[15] Para una discusión suplementaria, véase J. Woodward, «The Non-Identity Problem» [«El problema de la no identidad»], en *Ethics*, julio 1986.

Supongamos además que hemos planificado dos programas médicos, pero sólo hay fondos para uno; por lo que uno de ellos debe ser cancelado. En el primer programa, se analizaría a millones de mujeres durante el embarazo, y las que encontráramos que tienen la enfermedad J serían sometidas a tratamiento. En el segundo programa, se analizaría a millones de mujeres cuando tuvieran la intención de tratar de quedarse embarazadas. Las que encontráramos que tienen la enfermedad K serían advertidas para que pospusieran la concepción durante al menos dos meses, después de los cuales esta incurable enfermedad habrá desaparecido. Supongamos por último que podemos predecir que estos dos programas lograrían resultados en el mismo número de casos. Si hay análisis de embarazo, nacerían 1.000 niños normales en vez de discapacitados cada año. Si hay análisis de preconcepción, cada año nacerán 1.000 niños normales en vez de 1.000 niños diferentes discapacitados.

¿Valdrían la pena por igual estos dos programas? Tomemos nota cuidadosamente de cuál es la diferencia. Como resultado de cada programa, cada año 1.000 parejas tendrían un hijo normal en vez de discapacitado. Serían parejas diferentes en los dos programas. Pero como los números serían los mismos, los efectos sobre los padres y sobre otras personas serían moralmente equivalentes. Si hay una diferencia moral, sólo puede estar en los efectos sobre los niños.

Tomemos nota además de que, al juzgar estos efectos, no hay necesidad de tener ninguna opinión sobre el estatus moral del feto. Podemos suponer que transcurriría un año antes de que cada clase de análisis pudiera comenzar. Cuando elegimos entre los dos programas ninguno de los niños ha sido concebido todavía. Y todos los que son concebidos llegarán a ser adultos. Por eso estamos considerando los efectos, no sobre fetos presentes, sino sobre personas futuras. Asumamos además que la discapacidad en cuestión, aunque no sea insignificante, no es tan severa como para hacer dudoso que la vida de las personas afectadas sea digna de ser vivida. Aunque haber nacido pueda ir contra nuestros intereses, esto no ocurre con los que han nacido con esta discapacidad.

Como no podemos permitirnos los dos programas, ¿cuál deberíamos cancelar? Según una determinada descripción, los dos tendrían el mismo efecto. Supongamos que las enfermedades J y K son las únicas causas de esta discapacidad. La incidencia es ahora de 2.000 entre los que nacen cada año. Cada programa reduciría la incidencia a la mitad; la tasa caería a 1.000 cada año. La diferencia es esta. Si decidimos cancelar el análisis de embarazo, será verdadero de los que van a nacer después discapacitados que, si no fuera por nuestra decisión, se habrían curado. Nuestra decisión será peor para todas estas personas. Si en lugar de ello decidimos cancelar el análisis de preconcepción, más tarde habrá el mismo número de personas que nazca con esta discapacidad. Pero no sería verdadero de estas personas que, si no fuera por nuestra decisión, se habrían curado. Estas personas deben su existencia a nuestra decisión. Si no hubiéramos decidido cancelar el análisis de preconcepción, los padres de estos niños discapacitados no los habrían tenido *a ellos*. Habrían tenido, más tarde, hijos diferentes. Como las vidas de estos niños discapacitados valen la pena vivirse, nuestra decisión no será peor para ninguno de ellos.

¿Representa esto una diferencia moral? ¿O son los dos programas igualmente valiosos? ¿Todo lo que importa moralmente es cuántas vidas futuras serán vividas por personas normales, en vez de discapacitadas? ¿O acaso también importa el que estas vidas sean vividas por exactamente las mismas personas?

Deberíamos añadir un detalle al caso. Si decidimos cancelar el análisis de embarazo, los que nacen después discapacitados podrían saber que, si hubiéramos tomado una decisión diferente, se habrían curado. Este conocimiento podría hacer su discapacidad más difícil de soportar. Por ello deberíamos asumir que, aunque no se oculte de forma deliberada, estas personas no conocerían este hecho.

Con este detalle añadido, yo considero que los dos programas son igualmente valiosos. Sé de algunos que no aceptan esta afirmación, pero sé de más que sí la aceptan.

Mi reacción no es simplemente una intuición. Es el juicio al que llego razonando como sigue. Sea cual sea el programa cancelado, habrá después el mismo número de personas con esta discapacidad. Estas personas serían diferentes en los dos resultados que dependen de nuestra decisión. Y hay una afirmación que se aplica a sólo uno de estos dos grupos de personas discapacitadas. Aunque ellos no

conozcan este hecho, las personas de un grupo podrían haber sido curadas. Por tanto pregunto: «Si habrá personas con una discapacidad, el hecho de que estén discapacitadas es malo. ¿Sería *peor* en caso de que, sin saberlo ellas, su discapacidad pudiera haberse remediado?». Esto sería peor si este hecho hiciera a estas personas menos favorecidas de lo que lo son las personas cuya discapacidad *no* pudiera haber sido remediada. Pero este hecho no tiene este efecto. Si decidimos cancelar el análisis de embarazo, habrá un grupo de personas discapacitadas. Si decidimos cancelar el análisis de preconcepción, habrá un grupo diferente de personas discapacitadas. Las personas del primer grupo no serían menos favorecidas de lo que lo habrían sido las personas del segundo grupo. Por tanto, considero que estos dos resultados son moralmente equivalentes. Supuestos los detalles del caso, me parece irrelevante que uno de los grupos pero no el otro pudiera haber sido curado.

Este hecho *habría* sido relevante si curar a este grupo hubiera reducido la incidencia de la discapacidad. Pero, como sólo tenemos fondos para un programa, esto no es así. Si elegimos curar al primer grupo, después habrá el mismo número de personas con la discapacidad. Como curar al primer grupo no reduciría el número de los que estarán discapacitados, debemos elegir curar a este grupo sólo si tiene un derecho más sólido a ser curado. Y los miembros de este grupo no tienen un derecho más sólido. Si *pudiéramos* curar al segundo grupo, tendrían un derecho igual a ser curados. Si lo que elegimos fue curar al primer grupo, simplemente tendrían mejor suerte que el segundo grupo. Como simplemente tendrían mejor suerte, y no tienen un derecho más sólido a ser curados, yo no creo que debamos elegir curarlos. Como es también cierto que, si elegimos curarlos, esto no va a reducir el número de personas que estarán discapacitadas, yo concluyo que los dos programas son igualmente valiosos. Si el análisis de preconcepción lograse resultados en unos pocos casos más, yo consideraría que es el mejor programa [16].

[16] J. McMahan me ha sugerido que si la discapacidad afectase severamente la naturaleza de la vida de estas personas, puede que no esté claro que alguien con una discapacidad de por vida hubiera estado en mejor situación si hubiera

Esto se ajusta a mi reacción a nuestra elección de la Reducción. Pienso que sería malo que más adelante hubiese una gran bajada de la calidad de vida. Y pienso que no sería *peor* que las personas que vivan más adelante hubiesen existido también si hubiéramos elegido la Conservación. El mal efecto no sería peor si hubiese sido, de este modo, peor para alguna persona en particular. Al considerar los dos casos, yo acepto la Tesis de la No-Diferencia. Y, como yo, muchas otras personas.

He descrito dos casos en que yo y muchos otros, aceptamos la Tesis de la No-Diferencia. Si estamos en lo correcto al aceptarla, esto puede tener implicaciones teóricas de importancia. Lo cual depende de si pensamos que, si causamos que exista alguien que va a tener una vida digna de ser vivida, estamos con ello beneficiando a esa persona. Si pensamos así, aún no puedo formular las implicaciones de la Tesis de la No-Diferencia, puesto que dependerán de decisiones que todavía no he discutido. Pero supongamos que creemos que hacer que alguien exista no puede beneficiarle. Si esto es lo que pensamos, y aceptamos la Tesis de la No-Diferencia, las implicaciones son como sigue.

He sugerido que deberíamos apelar a

> C: Si en cada uno de dos resultados posibles viviese siempre el mismo número de personas, será peor que los que viven resulten menos favorecidos, o tengan una calidad de vida más baja, que los que habrían vivido.

Consideremos a continuación

nacido normal. Pues alguien puede dudar que, en el sentido relevante, estas dos vidas tan diferentes hubieran sido vividas por la misma persona. Y Adams (3) sugiere que, aunque tal persona hubiera estado en mejor situación, esto no tiene por qué implicar que fuese irracional para ella no lamentar su discapacidad. Si aceptamos cualquiera de las dos afirmaciones, el ejemplo no es lo que nos hace falta. Podemos eludir estas cuestiones suponiendo que la discapacidad afecta a estas personas sólo cuando son adultas. La discapacidad podría consistir, por ejemplo, en ser estériles.

La Tesis de las Personas Afectadas, o *T*: Será peor si las personas son afectadas para peor.

En las Elecciones de las Mismas Personas, C y T coinciden. Cuando consideramos estas elecciones, los que viven son los mismos en ambos resultados. Si estas personas resultan menos favorecidas, o tienen una calidad de vida más baja, son afectadas para peor, y viceversa [17]. Como C y T aquí coinciden, no supondrá ninguna diferencia a cuál de las dos apelamos.

Las dos afirmaciones entran en conflicto sólo en las Elecciones del Mismo Número. Son las que este capítulo ha discutido. Supongamos que aceptamos la Tesis de la No-Diferencia. Al considerar estas elecciones, apelaremos entonces a C *en vez de* a T. Si elegimos la Reducción, esto bajará el nivel de vida en el futuro lejano. De acuerdo con C, nuestra elección tiene un mal efecto. Pero, a causa de los hechos sobre la identidad, nuestra elección no será mala para nadie. T no implica que nuestra elección tenga un mal efecto. ¿Sería este efecto peor si *fuera* peor para personas concretas? Si apeláramos a T en vez de a C, nuestra respuesta sería Sí. Pero, como creemos en la Tesis de la No-Diferencia, contestaremos No. Creemos que aquí T da la respuesta equivocada. Y T da la respuesta equivocada en el caso de los Programas Médicos. C describe los efectos que creemos malos. Y pensamos que no supone ninguna diferencia moral el que estos efectos sean también malos de acuerdo con T. T traza distinciones morales donde, según nuestra opinión, no se deberían trazar.

En las Elecciones de las Mismas Personas, C y T coinciden. En las Elecciones del Mismo Número, donde estas afirmaciones entran en conflicto, aceptamos C antes que T. Cuando hacemos estas dos clases de elección, T, por tanto, no nos servirá de nada.

[17] Puede parecer que hay una excepción. Si mi vida vale la pena, matarme me afecta para peor, pero ¿me hace estar en peor situación o tener una calidad de vida más baja? Tal y como yo uso la frase, sí que tengo «una calidad de vida más baja». Esto es verdadero si mi vida va peor de lo que podría haber ido, o si lo que ocurre en mi vida es peor para mí. Las dos cosas son verdaderas cuando me matan, si tenía una vida que valía la pena.

Quedan las Elecciones de Diferente Número, que C no incluye. Aquí necesitaremos la Teoría X. Todavía no he discutido qué debería afirmar X. Pero podemos predecir lo siguiente: X implicará C en las Elecciones del Mismo Número.

Podemos predecir también que X tendrá la misma relación con T. En las Elecciones de las Mismas Personas, X y T coincidirán. Aquí no supondrá ninguna diferencia a cuál apelamos. Son las elecciones con las que la mayor parte de nuestro pensamiento moral está involucrado. Esto explica la verosimilitud de T. De esta parte de la moralidad, la que tiene que ver con la beneficencia, o el bienestar humano, se piensa usualmente en lo que denominaré términos de *las personas afectadas*. Apelamos a los intereses de las personas —a lo que es bueno o malo para las personas a las que nuestros actos afectan—. Aun después de que hayamos dado con la Teoría X, podríamos continuar apelando a T en la mayor parte de los casos, simplemente porque es más familiar. Pero en algunos casos X y T entran en conflicto. Pueden entrar en conflicto cuando hacemos Elecciones del Mismo y de Diferente Número. Y siempre que X y T entren en conflicto, apelaremos a X *antes que a* T. Pensaremos que, si un efecto es malo de acuerdo con X, no supone ninguna diferencia moral que sea también malo de acuerdo con T. Como antes, T traza una distinción moral donde, según nuestro parecer, no se debería trazar ninguna. T es como la afirmación de que es incorrecto esclavizar a los blancos, o negar el voto a los varones adultos. Por eso concluiremos que esta parte de la moralidad, la que tiene que ver con la beneficencia y el bienestar humano, no pude explicarse en términos de las personas afectadas. Sus principios fundamentales no tendrán que ver con si nuestros actos van a ser buenos o malos para estas personas a las que afectan. La Teoría X implicará que un efecto es malo si es malo para las personas. Pero esto no será *por lo que* este efecto sea malo.

Recordemos además que estas afirmaciones asumen que causar que se exista no puede beneficiar. Asunción que es defendible. Si asumimos tal cosa, estas afirmaciones mostrarán que muchas teorías morales necesitan revisarse, puesto que implican que tiene que

suponer una diferencia moral el que nuestros actos sean buenos o malos para las personas a las que afectan [18]. Y puede que necesitemos revisar nuestras creencias sobre ciertos casos comunes. Un ejemplo podría ser el aborto. Pero la mayor parte de nuestro pensamiento moral se mantendría inalterado. Muchas relaciones significativas se dan sólo entre personas concretas. Estas incluyen nuestras relaciones con aquellos a los que hemos hecho promesas, o debemos gratitud, o con nuestros padres, alumnos, pacientes, clientes, y (si somos políticos) con aquellos a los que representamos. Mis observaciones no se aplican a tales relaciones, ni a las obligaciones especiales a que dan pie. Mis observaciones se aplican sólo a nuestro Principio de Beneficencia: a nuestra razón moral general para beneficiar a otras personas, y protegerlas del mal.

[18] Un ejemplo es la convincente teoría presentada en Scanlon (3). Scanlon argumenta que la mejor explicación de la motivación moral no es la que dan los utilitaristas, que apelan a la filantropía universal. Nuestro motivo moral fundamental es, en cambio, «el deseo de poder justificar nuestras propias acciones ante los demás, recurriendo a razones que ellos no podrían rechazar de forma razonable». Scanlon esboza una atractiva teoría moral, construida sobre esta afirmación. Según esta teoría, un acto es incorrecto si va a afectar a alguien de un modo que no puede justificarse —si va a haber algún querellante cuya reclamación no puede ser contestada—. Según esta teoría, el marco de la moralidad tiene que ver esencialmente con las personas afectadas. Desafortunadamente, cuando elegimos una política como la Reducción Mayor, no habrá querellantes. Si somos de la opinión de que esto no representa ninguna diferencia moral, puesto que la objeción a nuestra elección sigue siendo igual de fuerte, pensamos que es irrelevante que no vaya a haber querellantes. El principio fundamental de la teoría de Scanlon traza una distinción allí donde, según nuestro parecer, no debería trazarse ninguna. Por eso necesita ser revisada la teoría de Scanlon.

Se aplican observaciones similares a muchas otras teorías. Así, Brandt (2) sugiere que a la frase «es moralmente incorrecto» le deberíamos asignar el significado descriptivo «estaría prohibido por cualquier código moral que todas las personas plenamente racionales tendiesen a apoyar, en preferencia a todos los demás o a ninguno en absoluto, para la sociedad del agente, si ellos esperasen pasar toda su vida en esa sociedad» (p. 194). Parece probable que, según el código elegido, un acto no sería incorrecto si no hubiese querellantes. Observaciones similares se aplican a Gert, a J. Narveson, *Morality and Utility* [*Moralidad y utilidad*], y a G. R. Grice, *The Grounds of Moral Judgement* [*Los fundamentos del juicio moral*], y pueden aplicarse a Mackie (2), Richards, Harman, Gauthier (4), Rawls, y otros.

Como mis observaciones se aplican nada más que a este principio, y habremos cambiado de opinión sólo en algunos casos, este cambio de opinión puede parecer carente de importancia. Pero no es así. Consideremos una vez más esta (demasiado grandiosa) analogía: en los casos corrientes podemos aceptar las Leyes de Newton. Pero no en todos los casos. Y ahora aceptamos una teoría diferente.

126. Causando catástrofes previsibles en el futuro más lejano

En esta sección, en vez de proseguir con la línea principal de mi argumento, discuto una cuestión menor. En un caso como el de la Reducción, no podemos resolver del todo el Problema de la No-Identidad apelando a los derechos de las personas. ¿Es esto también verdadero en una variante del caso en que nuestra elección causa una catástrofe? Como se trata de una cuestión menor, esta sección puede ignorarse, excepto por aquellos que no crean que la Reducción tiene un mal efecto. Consideremos

> *La Política Arriesgada*. Como comunidad, tenemos que elegir entre dos políticas energéticas. Las dos serían completamente seguras al menos durante tres siglos, pero una conllevaría riesgos en el futuro más lejano. Esta política supone el enterramiento de desechos nucleares en áreas en que, en los próximos siglos, no hay riesgo de terremoto. Pero como estos desechos seguirán radioactivos durante miles de años, habrá riesgos en el futuro lejano. Si elegimos esta Política Arriesgada, el nivel de vida será algo más alto en el próximo siglo. La elegimos. Como resultado, hay una catástrofe muchos siglos después. A causa de determinados cambios geológicos en la superficie terrestre, un terremoto libera la radiación, que mata a miles de personas. Aunque les mata esta catástrofe, estas personas habrán tenido vidas dignas de vivirse. Podemos suponer que esta radiación afecta sólo a personas que nacen después del escape, y que les hace contraer una enfermedad incurable que les mata, aproximadamente, a la edad de 40 años. Y la enfermedad no se manifiesta antes de que mate.

Nuestra elección entre las dos políticas influirá en los detalles de las vidas que se van a vivir posteriormente. Del modo explicado arriba, nuestra elección influirá, por tanto, en quién vivirá posteriormente. Tras muchos siglos no habría nadie viviendo en nuestra comunidad, que hubiera nacido fuera cual fuera la política que eligiéramos. Puesto que elegimos la Política Arriesgada, miles de personas mueren posteriormente. Pero si hubiéramos elegido la otra alternativa, la política segura, nunca habrían existido estas personas concretas. Habrían existido en su lugar personas diferentes. ¿Es peor para alguien nuestra elección de la Política Arriesgada?

Deberíamos preguntar: «si las personas viven vidas que vale la pena vivir, aunque mueran a consecuencia de una catástrofe, ¿es esto para ellas peor que si nunca hubieran existido?». Nuestra respuesta tiene que ser No. Aunque cause una catástrofe previsible, nuestra elección de la Política Arriesgada no será peor para nadie.

Hay quienes pueden decir que nuestra elección de la Reducción no tiene un mal efecto. Esto no puede decirse de la elección de la Política Arriesgada. Como esta elección causa una catástrofe, tiene claramente un efecto malo. Pero nuestra elección no será mala o peor para ninguna de las personas que vivan posteriormente. Este caso nos fuerza a rechazar la opinión de que una elección no puede tener un mal efecto si no va a ser mala para nadie.

En este caso, el Problema de la No-Identidad puede parecer más fácil de resolver. Aunque nuestra elección no sea peor para la gente alcanzada por la catástrofe, podría decirse que nosotros perjudicamos a esas personas. Y puede que aquí salga adelante la apelación a los derechos de las personas.

Podemos merecer que se nos culpe de dañar a otros, aunque no sea peor para ellos. Supongamos que conduzco mi coche temerariamente, y en el choque resultante provoco que pierdas una pierna. Un año después, estalla la guerra. Si no hubieras perdido la pierna, habrías sido reclutado obligatoriamente, y te habrían matado. Que yo haya conducido temerariamente, por tanto, te ha salvado la vida. Pero aun así soy culpable desde el punto de vista moral.

Este caso nos recuerda que, a la hora de asignar culpas, tenemos que considerar no los efectos reales sino los previsibles. Yo sé que

mi conducir temerario podría dañar a otros, pero yo no podía saber que de hecho te salvaría la vida. Esta distinción podría aplicarse a nuestra elección de la política arriesgada. Supongamos que sabemos que, si elegimos esta política, esto puede causar muchas muertes accidentales en el futuro lejano. Pero hemos pasado por alto el Problema de la No-Identidad. Creemos de forma equivocada que, sea cual sea la política que elijamos, posteriormente van a vivir las mismas personas. Por eso creemos que nuestra elección de la Política Arriesgada puede ir tremendamente en contra de los intereses de ciertas personas futuras. Si esto es lo que creemos, se puede criticar nuestra elección. Podemos merecer ser culpados por hacer lo que *creemos* puede ir tremendamente en contra de los intereses de otras personas. Esta crítica se mantiene en pie incluso si nuestra creencia es falsa —del mismo modo que tengo que ser también culpado aunque mi conducción temeraria de hecho te vaya a salvar la vida.

Supongamos que no podemos encontrar la Teoría X, o que X parece menos plausible que la objeción a hacer lo que puede ir muy en contra de los intereses de otras personas. Entonces puede ser mejor que les ocultemos el Problema de la No-Identidad a los que vayan a decidir si incrementamos nuestro uso de energía nuclear. Tal vez sea mejor que estas personas crean equivocadamente que semejante política puede, causando una catástrofe, ir muy en contra de los intereses de algunos de los que vivirán en el futuro lejano. Si tienen esta creencia falsa, puede ocurrir que lleguen con mayor probabilidad a las conclusiones correctas.

Nosotros hemos perdido esta falsa creencia. Nos damos cuenta de que, si elegimos la Política Arriesgada, nuestra elección *no* va a ser peor para las personas que la catástrofe mate más tarde. Nótese que esto no es una conjetura feliz. No es como predecir que, si causo que pierdas una pierna, esto más tarde te librará de morir en las trincheras. Sabemos que, si elegimos la Política Arriesgada, podemos provocar que mucha gente muera en el futuro lejano. Pero también sabemos que, si hubiéramos elegido la Política Segura, las personas que mueren nunca habrían nacido. Como las vidas de estas personas serán dignas de ser vividas, *sabemos* que nuestra elección no va a ser peor para ellas.

Si sabemos esto, no se nos puede comparar con el conductor temerario. ¿Cuál es la objeción a nuestra elección? ¿Puede ser incorrecto dañar a los demás cuando sabemos que nuestro acto no será peor para la gente dañada? Esto podría ser incorrecto si les pudiéramos haber pedido a estas personas su consentimiento, pero no hemos logrado hacerlo. Al no lograr pedirles a estas personas su consentimiento, infringimos su autonomía. Pero esto no puede ser la objeción a nuestra elección de la Política Arriesgada. Como no podríamos en absoluto comunicarnos con la gente que viva dentro de muchos siglos, no les podemos pedir su consentimiento.

Cuando no podemos pedir el consentimiento de alguien, deberíamos preguntarnos en vez de eso si esta persona lamentaría más tarde lo que estamos haciendo. ¿Lamentarían las personas que van a morir posteriormente nuestra elección de la Política Arriesgada? Supongamos que conocen todos los hechos. Desde una edad temprana saben que, a causa del escape de radiación, tienen una enfermedad incurable que les matará aproximadamente a la edad de 40 años. También saben que, si hubiéramos elegido la Política Segura, nunca habrían nacido. Lamentarían el hecho de que van a morir jóvenes. Pero, como sus vidas valen la pena de ser vividas, no lamentarían el hecho de que un día nacieron. Por tal razón no lamentarían nuestra elección de la Política Arriesgada.

¿Puede ser incorrecto dañar a los otros, cuando sabemos *no sólo* que si las personas dañadas conocieran nuestro acto no lo lamentarían, *sino además* que nuestro acto no será para ellas peor que ninguna otra cosa que pudiéramos haber hecho? ¿Cómo podríamos saber que, aunque dañamos a alguien, nuestro acto no será peor para esta persona? Hay como mínimo dos clases de casos:

(1) Aunque perjudicamos a alguien, podemos también saber que le estamos dando un beneficio que le compensa enteramente. Podríamos no saber esto a no ser que el beneficio tuviese claramente más peso que el daño. Pero, si esto es así, lo que hacemos será mejor para esta persona. En esta clase de casos, si tampoco infringimos su autonomía, puede que no haya ninguna objeción a nuestro acto. Puede que no haya objeción alguna a que perjudiquemos a alguien cuando sabemos no sólo que esta persona no lo va a lamentar, sino también que nuestro acto será claramente mejor para ella. En la ley inglesa, se consideró en una época la cirugía como daño corporal doloroso pero justificable. Como argumenté en la Sección 25, deberíamos revisar el uso corriente de la palabra «daño». Si lo que hacemos no va a ser peor para alguna otra persona, o incluso va a ser mejor para ella, no la estamos dañando en ningún sentido moralmente relevante.

Si asumimos que causar que se exista puede beneficiar, nuestra elección de la Política Arriesgada es, en sus efectos sobre los muertos, como el caso de la cirugía. Aunque nuestra elección provoca que esa gente muera, como también causa que exista con una vida digna de ser vivida les da un beneficio que pesa más que este daño. Lo cual sugiere que la objeción a nuestra elección no puede ser que daña a estas personas.

Podemos asumir en cambio que causar que se exista no puede beneficiar. Según esta suposición, nuestra elección de la Política Arriesgada no da a las personas que mata un beneficio que les compense completamente. Nuestra elección no es *mejor* para ellas. Simplemente *no es peor* para ellas.

(2) Hay otra clase de casos en que podemos saber que, aunque dañamos a alguien según el uso corriente de «dañar», esto no va a ser peor para esta persona. Son los casos que conllevan sobredeterminación. En ellos, sabemos que, si no dañamos a nadie, la persona en cuestión será dañada como mínimo en la misma medida de alguna otra manera. Supongamos que alguien está atrapado en unas ruinas, a punto de morir abrasado. Nos pide que le disparemos para no morir con dolor. Si matamos a esta persona no la estamos dañando, en un sentido moralmente relevante.

Semejante caso no puede demostrar que no haya objeción alguna a nuestra elección de la Política Arriesgada, puesto que no es similar de una forma relevante. Si no ocurriera la catástrofe, la gente muerta habría vivido muchos años más. Hay una razón muy diferente por la que nuestra elección de la política arriesgada no es peor para estas personas.

¿Podría haber un caso en que matemos a una persona existente, sabiendo lo que sabemos cuando elegimos la Política Arriesgada?

Tenemos que saber (a) que esta persona se enterará del hecho de que hemos realizado algo que causará que muera, pero no lo lamentará. Y tenemos que saber (b) que, aunque esta persona habría vivido en caso contrario una vida normal durante muchos años más, causar que muera no será ni mejor ni peor para ella. ((b) es lo que sabemos de los efectos de nuestra elección de la Política Arriesgada, si asumimos que, al hacer lo que es una parte necesaria de la causa de la existencia de la gente muerta por la catástrofe, no podemos estar beneficiando a estas personas).

Supongamos que matamos a una persona existente, una persona que, si no, habría vivido una vida normal durante muchos años más. En semejante caso, no podríamos *saber* que (b) es verdadero. Aunque vivir durante estos muchos años no sería ni mejor ni peor para esta persona, esto nunca se podría prever. No puede haber un caso en que matemos a una persona existente sabiendo lo que sabemos cuando elegimos la Política Arriesgada. Un caso que sea similar de una forma relevante tiene que implicar causar que alguien muera, alguien que, si nosotros hubiéramos obrado de otro modo, nunca habría existido.

Comparemos estos dos casos:

La Elección de Jane. Jane tiene una enfermedad congénita que la matará sin dolor aproximadamente a la edad de 40 años. Se trata de una enfermedad que no tiene efectos antes de matar. Jane sabe que, si tiene un hijo, tendrá la misma enfermedad. Supongamos que también puede asumir lo siguiente. Como ella, su hijo tendría una vida digna de ser vivida. No hay niños que necesiten ser adoptados pero que no lo hayan sido. Dado el tamaño de la población mundial cuando ocurre este caso (quizás en algún siglo futuro), si Jane tiene un hijo, esto no será peor para otras personas. Y si no lo tiene, no podrá criar a ningún niño. No puede convencer a otra persona de que tenga un hijo extra, a quien ella criaría. (Estas asunciones nos dan la cuestión relevante.) Conociendo estos hechos, Jane elige tener un hijo.

La elección de Ruth. La situación de Ruth es igual que la de Jane, con una diferencia. Su enfermedad congénita, a diferencia de la de Jane, sólo mata a los varones. Si Ruth se costeara la nueva técnica de

fecundación *in vitro*, estaría segura de tener una hija a quien la enfermedad no mataría. Pero decide ahorrarse el gasto y correr el riesgo. Desafortunadamente, tiene un hijo, cuya enfermedad heredada le matará más o menos a los 40.

¿Hay una objeción moral a la elección de Jane? Dados los supuestos del caso, la objeción tendría que apelar al efecto en el hijo de Jane. Su elección no fue peor para ese niño. ¿Hay una objeción a su elección que apele a los derechos del niño? Supongamos que tenemos la creencia de que cada persona tiene el derecho de vivir una vida completa. Jane sabe que, si tiene un hijo, su derecho a una vida completa no va a poder realizarse de ningún modo. Lo cual puede implicar que Jane no vulnera este derecho. Pero la objeción se podría reformular. Podría decirse: «Es incorrecto causar que exista alguien con un derecho que no puede realizarse. Por eso actúa Jane incorrectamente».

¿Es esta una buena objeción? Si fuera yo el hijo de Jane, mi opinión sería como la del hombre que escribió a *The Times*. Yo lamentaría el hecho de ir a morir joven. Pero, como mi vida vale la pena de ser vivida, no lamentaría que mi madre me hiciera existir. Y negaría que su acto fuese incorrecto a causa de lo que me hizo a mí. Si me dijeran que *fue* incorrecto porque me hizo existir con un derecho que no se puede realizar, yo *renunciaría* a ese derecho.

Si el hijo de Jane renunciara a su derecho, eso podría socavar esta objeción a su elección. Pero, aunque *yo* renunciaría a este derecho, no puedo estar seguro de que, en todos los casos como este, vaya a ser esto lo que el niño haga. Si el hijo de Jane no renuncia a su derecho, una apelación al mismo puede quizás proporcionar una objeción a su elección.

Volvamos ahora a la elección de Ruth. Claramente, hay una objeción mayor a *esta* elección. Esto es porque Ruth tiene una alternativa diferente. Si Jane no tiene un hijo, no podrá criar a un niño, y se vivirá una vida menos. La alternativa de Ruth es pagar por la técnica que le dará un niño diferente, al que la enfermedad no matará. Ella elige ahorrarse el gasto, sabiendo que tiene una de entre dos probabilidades de que a su hijo lo mate la enfermedad.

Aunque haya una objeción a la elección de Jane, hay una objeción mayor a la de Ruth. Esta objeción no puede apelar sólo a los

efectos sobre el hijo real de Ruth, puesto que son iguales que los efectos de la elección de Jane sobre el hijo de Jane. La objeción a la elección de Ruth tiene que apelar en parte al posible efecto sobre el niño diferente que, pagando por la nueva técnica, ella podría haber tenido. La apelación a este efecto no es una apelación a los derechos de nadie.

Volvamos ahora a nuestra elección de la Política Arriesgada. Si la elegimos, esto puede causar que existan personas que morirán en una catástrofe. Sabemos que nuestra elección no sería peor para ellas. Pero, si tiene fuerza la objeción a la elección de Jane, podría aplicarse esta objeción a nuestra elección. Eligiendo la Política Arriesgada, podemos causar que existan personas cuyo derecho a una vida completa no puede realizarse.

La apelación a los derechos de estas personas puede aportar alguna objeción a nuestra elección. Pero no puede aportar la objeción completa. Nuestra elección es, en un aspecto, diferente de la de Jane. Su alternativa era no tener ningún hijo. Nuestra alternativa es como la de Ruth. Si hubiéramos elegido la Política Segura, habríamos tenido diferentes descendientes, ninguno de los cuales hubiera muerto por causa del escape radioactivo.

La objeción a la elección de Ruth no puede apelar sólo al derecho de su hijo a una vida completa. Lo mismo es verdadero, por tanto, de la objeción a nuestra elección de la Política Arriesgada. Esta objeción tiene que apelar en parte a los efectos en las posibles personas que habrían vivido si hubiéramos elegido de forma diferente. Como antes, la apelación a los derechos no puede resolver del todo el Problema de la No-Identidad. También tenemos que apelar a una tesis como C, que compara dos conjuntos diferentes de vidas posibles.

Puede objetarse: «Cuando Ruth concibe a su hijo, él hereda la enfermedad que le negará una vida completa. Puesto que la enfermedad de este niño se hereda de esta forma, no puede decirse que la elección de Ruth mate a su hijo. Si elegimos la Política Arriesgada, las conexiones causales serán menos estrechas. Puesto que las conexiones son menos estrechas, nuestra elección mata a las personas que más tarde mueren de los efectos de la radiación liberada. Que matamos a estas personas es la objeción completa a nuestra elección».

Encuentro discutible esta objeción. ¿Por qué hay una objeción mayor a nuestra elección porque las conexiones causales sean menos estrechas? La objeción puede ser correcta en lo que afirma acerca de nuestro uso corriente de «matar». Pero, como defendí en la Sección 25, este uso es moralmente irrelevante. Como ese argumento puede no convencer, añado

> *La Arriesgada Cura de la Esterilidad.* Ann no puede tener hijos como no se someta a determinado tratamiento. Si se somete al tratamiento, tendrá un hijo varón sano. Pero existe el riesgo de que este tratamiento le vaya a generar una rara enfermedad. Una enfermedad con las siguientes características. Es indetectable, y no ataca a las mujeres, pero puede contagiar a los parientes más cercanos de uno. Por consiguiente lo que sigue es verdadero: si Ann se somete a este tratamiento y tiene un hijo sano, hay una probabilidad entre dos de que más tarde le contagie de un modo que acabe por matarle cuando tenga aproximadamente cuarenta años. Ann elige someterse a este tratamiento, y, en efecto, después contagia a su hijo de esta enfermedad mortal.

Según la objeción expuesta arriba, hay una poderosa objeción a la elección de Ann, que no se aplica a la de Ruth. Puesto que las conexiones causales son menos directas, la elección de Ann mata a su hijo. Y ella sabía que la probabilidad de que su elección tuviera este efecto era de una entre dos. Ruth sabe que existe la misma probabilidad de que su hijo vaya a morir aproximadamente a la edad de 40 años. Pero, como las conexiones causales son tan directas, su elección no mata a su hijo. De acuerdo con esta objeción, esta diferencia tiene gran relevancia moral.

Esto no es plausible. Tanto Ruth como Ann saben que, si obran de determinada forma, hay una probabilidad entre dos de que tengan hijos que morirán por enfermedad alrededor de los cuarenta. La historia causal es diferente. Pero esto no hace la elección de Ann moralmente peor. Creo que este ejemplo demuestra que deberíamos rechazar esta última objeción.

El objetor podría decir: «Yo niego que, eligiendo someterse a la Cura Arriesgada, Ann mate a su hijo». Pero, si el objetor niega esto, no puede afirmar que, eligiendo la Política Arriesgada, nosotros matemos a ciertas personas en el futuro lejano. Las conexiones causales adoptan la misma forma. Cada elección produce un efecto secundario que más tarde mata a personas que deben su existencia a esta elección.

Si esta objeción fracasa, como creo, mi afirmación anterior está justificada. Es moralmente significativo que, si elegimos la Política Arriesgada, nuestra elección es como la de Ruth más bien que como la de Jane. Es moralmente significativo que, si hubiéramos elegido de otro modo, habrían vivido diferentes personas a las que no se habría matado. Como esto es así, la objeción a nuestra elección no puede apelar sólo a los derechos de los que realmente viven después. Tiene también que apelar a una afirmación como C, que compara diferentes conjuntos de vidas posibles. Como dije antes, la apelación a los derechos no puede resolver del todo el Problema de la No-Identidad.

127. Conclusiones

Ahora resumiré mis afirmaciones. Es en efecto verdadero de cada uno que, si no hubiera sido concebido en el espacio de un mes alrededor del momento en que fue concebido, nunca habría existido. Como esto es cierto, podemos afectar fácilmente a las identidades de las personas futuras, o a *quiénes* sean las personas que vivirán más adelante. Si una elección entre dos políticas sociales va a afectar al nivel de vida o a la calidad de vida durante un siglo, afectará a los detalles de todas las vidas que, en nuestra comunidad, se vivan más adelante. Como consecuencia, algunos de los que vivan más adelante deberán su existencia a nuestra elección de una de estas dos políticas. Tras uno o dos siglos, esto será verdadero de todo el mundo en nuestra comunidad.

Este hecho provoca un problema. Una de estas dos políticas puede, en el futuro distante, causar una gran disminución de la calidad de vida. Tal sería el efecto de la política que llamo Reducción. Este efecto es malo, y aporta una razón moral para no elegir la Reducción. Pero, debido al hecho recién mencionado, nuestra elección de la Reducción no será peor para nadie. Algunos piensan que una elección no puede tener malos efectos si no va a ser peor para nadie. El caso de la Reducción demuestra que debemos rechazar esta opinión. Y esto lo demuestra con más fuerza el caso de la Política Arriesgada. Un efecto de elegir esta política es una catástrofe que mata a miles de personas. Este efecto es evidentemente malo, aunque nuestra elección no será peor para nadie.

Como estas dos elecciones no serán peores para nadie, necesitamos explicar por qué tenemos una razón moral para no hacerlas. Este problema surge porque, en los diferentes resultados, existirían personas diferentes. Por eso lo llamo el Problema de la No-Identidad.

Pregunté si podemos resolver este problema apelando a los derechos de las personas. Argumenté que, incluso en el caso de la Política Arriesgada, la objeción a nuestra elección no puede apelar solamente a los derechos de las personas. La objeción tiene que apelar en parte a una afirmación como C, que compara vidas posibles diferentes. Y no podemos apelar con verosimilitud a los derechos a la hora de explicar la objeción a nuestra elección de la Reducción Menor. Incluso después de la gran disminución de la calidad de vida, los que vivirán estarán en una situación mucho mejor que la nuestra ahora. No se puede decir que estas personas tengan el *derecho* a la calidad de vida aun más elevada que habrían disfrutado personas diferentes si nosotros hubiéramos elegido la Conservación. Si eliminamos con la imaginación el Problema de la No-Identidad, la objeción a nuestra elección apelaría a nuestro Principio de Beneficencia. Para resolver el Problema de la No-Identidad, tenemos que revisar este principio.

Un principio revisado es C, la Tesis de la Calidad del Mismo Número. De acuerdo con C, si en cada uno de dos resultados hubiera el mismo número de personas, sería peor si los que viven resultan menos favorecidos, o tienen una calidad de vida más baja, que los que habrían vivido. Necesitamos un principio más amplio para

incluir casos en que, en los diferentes resultados, habría números diferentes de personas. Este principio que necesitamos lo llamo Teoría X. Sólo X resolverá del todo el Problema de la No-Identidad.

¿Supone el hecho de la no-identidad una diferencia moral? Cuando vemos que nuestra elección de la Reducción no será peor para nadie, podemos pensar que hay una objeción menor a nuestra elección. Pero creo que la objeción es igual de sólida. Y tengo una creencia similar cuando comparo los efectos de los dos Programas Médicos. A esta creencia la llamo la Tesis de la No-Diferencia. Aunque sé de personas que no la aceptan, sé de más que sí la aceptan. Si la aceptamos, y pensamos que causar que se exista no puede beneficiar, esto tiene amplias implicaciones teóricas. Podemos predecir que la Teoría X no adoptará una forma de personas afectadas. La mejor teoría de la beneficencia no apelará a lo que es bueno o malo para las personas a las que nuestros actos afectan.

En lo que sigue, trataré de encontrar la Teoría X. Como he dicho, este intento hará surgir algunas cuestiones enigmáticas.

17

LA CONCLUSIÓN REPUGNANTE

¿Cuántas personas debería haber? ¿Puede haber *superpoblación*: demasiadas personas? Necesitamos una respuesta a estas preguntas que resuelva también el Problema de la No-Identidad.

Más adelante preguntaré cuántas personas debería haber *en cualquier momento*. En una teoría moral completa, no podemos evitar esta impresionante pregunta. Y nuestra respuesta puede tener implicaciones prácticas. Puede, por ejemplo, afectar a lo que pensamos sobre las armas nucleares.

En la mayor parte de lo que sigue, discuto una pregunta más pequeña. ¿Cuántas personas debería haber, en algún país o en el mundo, durante un período determinado? ¿Cuándo habría demasiadas personas viviendo?

128. ¿Es mejor que vivan más personas?

Consideremos

El Niño Feliz. Una pareja trata de decidir si tener otro hijo. Puede suponerse que, si lo tuvieran, le querrían y su vida sería perfecta-

mente digna de vivirse. Dado el tamaño de la población mundial cuando este caso ocurre (quizás en algún siglo futuro), esta pareja puede suponer además que tener otro hijo no sería peor para otras personas, una vez considerados todos los factores. Cuando se ponen a considerar cuáles serán los efectos en ellos mismos, tienen varias razones para tener otro hijo; pero también tienen varias razones opuestas, como por ejemplo el efecto en sus carreras. Como muchos otros, esta pareja no puede decidir entre estos dos conjuntos de razones en conflicto. Piensan que, si tuvieran el hijo, no iba a ser ni mejor ni peor para ellos.

¿Tiene esta pareja una razón moral para tener este hijo? ¿Sería mejor que se viviera una vida más, digna de vivirse? Algunos responden Sí. Si esta pareja tiene una razón moral para tener a este hijo, y no puede decidir entre sus otras razones, su razón moral puede inclinar la balanza. Tal vez deban tener a este niño.

Otras personas adoptan una perspectiva diferente. Piensan que *no* existe ninguna razón moral para tenerlo. Según su opinión, *no* es mejor que se viva una vida extra digna de vivirse.

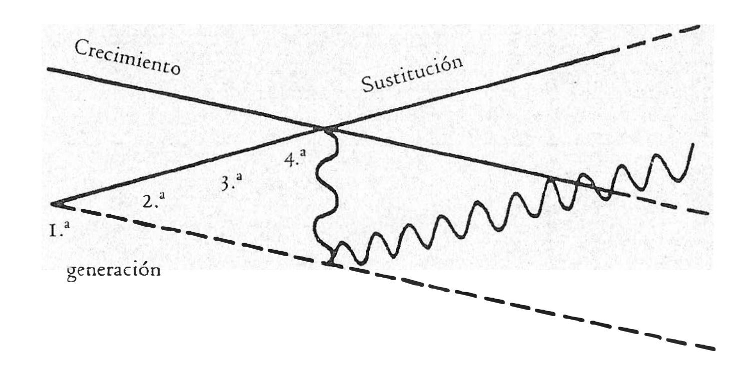
129. Los efectos del crecimiento demográfico en las personas existentes

La pareja de mi ejemplo da por sentado que la existencia de un niño más no sería peor para otras personas, una vez considerados todos los factores. En muchos países, en muchas épocas, esto ha sido verdadero. Pero en otras épocas no. En estas, si hubiera habido más personas, la gente habría salido perjudicada. Esto es lo que ocurre ahora en muchos países, en los que, si la población crece, la calidad de vida será más baja de lo que sería si la población no creciera. Estos son los casos que discutiré.

En ellos, el crecimiento demográfico afectaría a las personas existentes. Cuando el crecimiento demográfico baja la calidad de vida, podríamos pensar que tiene que ir en contra de los intereses de las personas existentes. Pero esto no es siempre cierto. Aunque baje la calidad de vida, el crecimiento demográfico puede ser *mejor* para las personas existentes.

Lo siguiente puede ser cierto en algún país durante una época. Si la población crece a un cierto ritmo, esto tendrá buenos efectos transitorios, y malos efectos acumulativos. Los malos efectos podrían consistir en la disminución continua de la cuota por persona de los recursos disponibles. Los buenos efectos transitorios podrían consistir en el funcionamiento de la economía de este país. Mientras que el rápido crecimiento demográfico puede ser malo para la economía, un ritmo lento de crecimiento puede ser mejor para la economía que ningún crecimiento en absoluto. Hay varias razones técnicas por las que esto puede ser verdadero para determinados sistemas económicos. Una analogía remota es el hecho de que, cuando estamos dando una curva con el coche, podemos llevar el volante más fácilmente si aceleramos. Y el crecimiento demográfico puede tener otros buenos efectos transitorios, de naturaleza no económica: podrían ser simplemente que las familias más grandes tienden a ser de algún modo más felices, o que muchas personas prefieren tener más niños.

Robertson escribe que, si determinado ritmo de crecimiento demográfico tuviera estas dos clases de efectos, «lo que sería... más conveniente es una población que esté siempre creciendo, pero que nunca llegue a ser demasiado grande» [19]. Únicamente podríamos alcanzar este ideal en el País de las Maravillas de Alicia. En el mundo real, las alternativas se pueden mostrar de la manera siguiente:



[19] Robertson, p. 460.

A esto lo llamo el *Caso de la Escalera Mecánica de Bajada*. Una línea muestra la calidad de vida que resultaría de mantener una población estable, o lo que yo llamo *Sustitución*. La línea de rayas muestra los malos efectos acumulativos de un determinado ritmo de crecimiento demográfico, muestra lo que sería la calidad de vida si no hubiera buenos efectos transitorios. La otra línea continua muestra los efectos combinados de este ritmo de crecimiento. Esto es lo que sería la calidad de vida, contando con que hubiera buenos efectos transitorios.

Como indica el diagrama, si hay Crecimiento en vez de Sustitución, la calidad de vida será más elevada para las tres primeras generaciones, pero después será cada vez más baja. Muchos creen que sería malo que el crecimiento demográfico bajase la calidad de vida. Para estas personas, el Caso de la Escalera Mecánica de Bajada es especialmente deprimente. Parece probable que el resultado real fuese el de Crecimiento. Puede ser mejor para la mayoría de las parejas tener más de dos hijos; y puede ser lo que la mayoría de las parejas quieran hacer. Además, como el Crecimiento sería mejor para las personas existentes y para las dos generaciones siguientes, es improbable que la comunidad se vaya a decidir por una política que causase un cambio del Crecimiento a la Sustitución.

Podría decirse: «Esto no es así. La mejor política sería el Crecimiento para las tres primeras generaciones, seguido por la Sustitución. La comunidad debería cambiar a la Sustitución una vez que el Crecimiento comience a producir una calidad de vida más baja. En este punto el Crecimiento deja de ser mejor para las personas existentes. Como esto es así, podemos esperar que la comunidad cambie a la Sustitución».

Esto es un error. Consideremos las alternativas a las que se enfrenta la cuarta generación. Si ha habido Crecimiento durante las tres primeras generaciones, ¿por qué la calidad de vida es aún tan alta como lo habría sido si hubiera habido Sustitución? Sólo por causa de los buenos efectos transitorios del Crecimiento. Si la cuarta generación cambia a la Sustitución, perderá estos buenos efectos. Su calidad de vida caerá al punto que se halla debajo en sentido vertical, sobre la línea de rayas. Y si continúa habiendo Sustitución, la calidad de vida será, para las próximas tres generaciones, más baja

de lo que habría sido si hubiera habido crecimiento demográfico. Estos dos efectos los muestran las líneas onduladas. Como indican estas, la cuarta generación tiene, a niveles más bajos, las mismas alternativas que la primera. *Toda* generación tiene estas alternativas.

Estas seguirán siendo las alternativas mientras el Crecimiento tenga estos dos efectos: los malos efectos acumulativos y los buenos efectos transitorios. Siempre y cuando esto sea así, comparado con la Sustitución, el Crecimiento sería siempre mejor tanto para las personas existentes como para las dos generaciones siguientes. Por consiguiente es probable que toda generación elija el Crecimiento. Como resultado, la calidad de vida continuará bajando. Si pensamos que esto es un mal efecto, el Caso de la Escalera Mecánica de Bajada es, como dije, especialmente deprimente. Es un Dilema del Prisionero Intergeneracional, de una clase en la que es muy poco probable que los que están implicados lleguen a una solución [19b].

[19b] Merece la pena subrayar que este caso es un Dilema Intertemporal Cada Uno-Nosotros, con dos rasgos especiales. Como afecta a generaciones diferentes, las personas afectadas no pueden comunicarse para alcanzar alguna clase de solución política, o algún acuerdo condicional conjunto. Y este es un dilema de esa clase especialmente intratable *que incluye a intrusos*.

Consideremos el *Dilema del Auditorio*. Si la primera fila se pone de pie, mejorará su visión del absorbente espectáculo que se desarrolla en el escenario. Si merece la pena ponerse de pie para conseguir esta visión mejor, será mejor para la primera fila ponerse de pie. Pero esto le impediría la visión a la segunda fila, de manera que esta necesitaría ponerse de pie para recuperar la visión que tenía cuando todos estaban sentados. Como ahora estaría de pie pero no habría mejorado su visión, este resultado sería peor para la segunda fila. Observaciones similares se aplican a todas las demás filas.

Este caso difiere de un Dilema Cada Uno-Nosotros corriente. Hay dos actos: A (más altruista), E (más egoísta). En un Dilema corriente, será mejor para cada uno si hace E, hagan los demás lo que hagan; pero que todos hagan E será peor para cada uno que si todos hacen A. En el Dilema del Auditorio, hay una diferencia pequeña pero fatídica. Será mejor para cada fila si se pone de pie en vez de permanecer sentada, pero si todas se ponen de pie en vez de permanecer sentadas eso *no* será peor para *todas* las filas. Será peor para todas las filas *excepto la primera*. La primera fila es *el intruso* en este Dilema.

Ya que contienen intrusos, tales Dilemas son especialmente intratables. El patrón de actos que es peor para cada uno de los demás es mejor para los intrusos.

Podemos suponer que, al juzgar los efectos de nuestros actos, basta con considerar los intereses de todas las personas que vayan a

De manera que sería peor para los intrusos que ayudaran a concertar una solución política, o se sumaran a un acuerdo condicional. Y lo que los intrusos hacen puede iniciar una viciosa reacción en cadena, que haga peor para cada uno sumarse a semejante acuerdo. De este modo, en el Dilema del Auditorio, será peor para la primera fila que todos se sienten, en lugar de estar de pie. Por tanto, será peor para esta fila sumarse al acuerdo de que todos deberían sentarse. Por tanto puede quedarse de pie. Una vez que la primera fila se ha puesto en pie, será peor para la segunda fila sumarse al acuerdo de que todos salvo la primera fila deberían quedarse sentados. Por tanto, puede ponerse de pie. Entonces será peor para la tercera fila sumarse al acuerdo de que todos salvo las dos primeras filas deberían quedarse sentados. Por tanto puede ponerse de pie. Observaciones semejantes se aplican a cada fila. El resultado final puede ser que todas las filas se pongan en pie en lugar de quedarse sentadas. Lo cual es peor para cada una de las filas excepto para la primera. La presencia de la primera fila, el intruso, impide aquí el logro del acuerdo condicional conjunto. Y la misma reacción en cadena puede impedir el logro de una solución política. Este rasgo especial hace que Dilemas como este cuenten con menos probabilidades de ser resueltos.

Además de ser trivial, el Dilema del Auditorio carece del otro rasgo deprimente. Afecta a contemporáneos. Esto hace que cuente con más probabilidades de ser resuelto. Las otras filas pueden recurrir a amenazas para mantener sentada a la primera fila. O bien la primera fila podría sentarse simplemente porque esperase quejas de las otras filas.

Un Dilema intergeneracional *no* afecta a contemporáneos. Esto lo hace más difícil de resolver. En estos Dilemas, si todos más bien que ninguno dejan de darse ciertas clases de prioridad a sí mismos, esto será mejor para todas las generaciones, *excepto la primera*. Las diferentes generaciones no se pueden comunicar para llegar a un acuerdo condicional conjunto. Ni tampoco pueden las generaciones anteriores ser disuadidas por las amenazas de las generaciones posteriores. Por eso representa un problema mayor que este Dilema contenga intrusos. En un Dilema intergeneracional —que no tiene por qué conllevar crecimiento demográfico— la generación existente siempre está en la posición de una fila después de que las filas anteriores se hayan puesto en pie. Ya ha sufrido por causa de la conducta de las generaciones anteriores. Y esta conducta anterior no puede alterarse ahora por ninguna solución moral o política. Como las cosas son así, sería peor para la generación existente tomar parte en tal solución. No sería movida por una renuencia a «gorronear», puesto que no puede beneficiarse de esta solución. Perdería con su propio acto, y no ganaría nada a cambio. Es así menos probable que tome parte en una solución. El mismo razonamiento se aplicará entonces a la generación siguiente y a todas las sucesivas.

vivir alguna vez. En el Caso de la Escalera Mecánica de Bajada, un determinado ritmo de crecimiento demográfico irá siempre a favor de los intereses de las personas existentes y de sus hijos. Aunque cause una continua disminución de la calidad de vida, este ritmo de crecimiento no irá en contra de los intereses de los que viven más de tres generaciones después. Esto es así porque, de la manera explicada en la Sección 123, estas personas deberán su existencia a este ritmo de crecimiento. Si sólo apelamos a los intereses de todas las personas que vayan a vivir alguna vez, tenemos que afirmar que sería *mejor* que hubiera este ritmo de crecimiento, a pesar de la continua disminución de la calidad de vida. Si queremos evitar esta conclusión, este es otro caso en que tenemos que apelar a un principio de una clase diferente, uno que no se plantea en términos de *personas afectadas*, o de los intereses de las personas.

Para los que deploran una disminución de la calidad de vida, este es el caso más deprimente. Afortunadamente, es sólo uno entre varios posibles. Hay dos modos en que, cuando baja la calidad de vida, el crecimiento demográfico puede ser *peor* para las personas existentes.

En algunas comunidades, será peor para la mayoría de las parejas tener más de dos hijos. Estos son los casos en que es poco probable que haya crecimiento demográfico.

En otras comunidades, como expliqué en el capítulo 3, la mayoría de las parejas se enfrenta a un Dilema del Prisionero. En estas comunidades, será mejor para cada una de muchas personas que él o ella tenga más de dos hijos, hagan lo que hagan las demás personas. Pero si todos tienen más de dos hijos eso será para cada uno peor que si tienen menos. Si estas personas llegaran a ver que esto era verdadero, podrían lograr lo que llamo una solución política. Aunque cada uno preferiría tener más hijos, cada uno podría también preferir que nadie tuviera más hijos antes que todos los tuvieran. Podría adoptarse democráticamente un sistema de recompensas y de multas, con el objetivo de frenar el crecimiento demográfico. Y aunque se impusiera de modo no democrático, semejante sistema podría ser bien recibido por todas estas personas. Otra solución nos la daría la esterilización reversible tras el nacimiento

del segundo hijo. Es una solución mejor, porque no habría que imponer multas. Una vez más, si comprendieran los hechos, todas estas personas podrían darle la bienvenida.

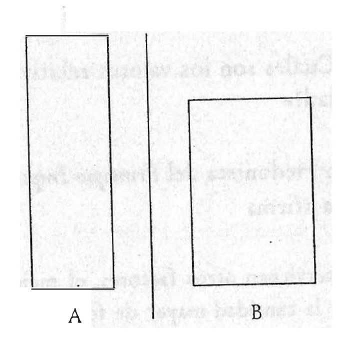
He descrito tres clases de casos. Y hay otros casos posibles. Algunos serían mezclas de estos tres, pero otros serían diferentes en otros aspectos.

130. Superpoblación

Cuando el crecimiento demográfico baja la calidad de vida, los efectos en las personas existentes pueden ser buenos, malos, o ni lo uno ni lo otro. Estos efectos no hacen surgir nuevas cuestiones morales. Pero otros efectos sí las hacen surgir.

Estas cuestiones surgen con la mayor claridad cuando comparamos los resultados que serían producidos, en el futuro lejano, por diferentes ritmos de crecimiento demográfico. Si hay un crecimiento más rápido, después habrá más personas que resultarán menos favorecidas. Como antes, «menos favorecidas» puede referirse o bien al nivel de felicidad, o a la calidad de vida, o a la cuota por persona de los recursos. Deberíamos suponer que, en mis ejemplos, estos tres correlacionan, subiendo y disminuyendo juntos.

Comparemos los resultados de dos ritmos de crecimiento demográfico, después de uno o dos siglos. Como he explicado, no habría nadie que existiera en ambos resultados. Dos resultados así se muestran abajo



La anchura de cada bloque indica el número de personas vivas, la altura su calidad de vida. Con esto me refiero a su calidad de vida durante una época. En tal época, habría algún cambio en la población. Pero en aras de la simplicidad podemos ignorar este hecho. Por la misma razón, podemos suponer que en estos resultados no hay desigualdad ni social ni natural; nadie resulta menos favorecido que nadie. Esto nunca sería verdadero de hecho. Pero no puede distorsionar nuestro razonamiento sobre las preguntas que haré el que nos imaginemos que sería verdadero. Y esto hace que mis preguntas adopten una forma más clara.

En B hay dos veces más personas vivas que en A, y todas ellas están en peor situación que todas las que hay en A. Pero las vidas de las de B, comparadas con las de A, valen la pena de vivirse más de la mitad de lo que lo valen las otras. Esta afirmación no supone que, como sugiere mi diagrama, estos juicios pudieran en principio ser precisos. Pienso que se da sólo una posibilidad de comparación a grandes rasgos o parcial. Lo que supone mi afirmación es que un desplazamiento del nivel en A al nivel en B sería una disminución en la calidad de vida, pero que haría falta mucho más que otra disminución igual de grande para que las vidas de las personas dejaran de ser dignas de vivirse.

Hay diversos modos en que, con el doble de población, la calidad de vida podría ser más baja. Podría haber peores viviendas, escuelas atestadas de gente, más polución, menos belleza natural, y una renta media de algún modo más baja. Si estas son las maneras en que la calidad de vida sería más baja, podemos suponer con verosimilitud que haría falta mucho más que otra disminución similar antes de que la vida dejase de valer la pena.

Salvo por la ausencia de desigualdad, estos dos resultados podrían ser las alternativas reales en algún país, o para la humanidad, dados dos ritmos de crecimiento demográfico durante muchos años. ¿Cuál sería el mejor resultado? Por «mejor» no quiero decir «moralmente mejor» en el uso más común de esta expresión. Este se aplica sólo a personas o a actos. Pero uno de los dos resultados puede ser mejor en otro sentido, que tiene relevancia moral. Sería mejor, en este sentido, que menos personas sufriesen de enferme-

dades graves, o que no hubiera ocurrido el terremoto de Lisboa. Y nosotros podemos, evidentemente, hacer afirmaciones como estas sobre resultados que involucran a diferentes poblaciones posibles. Supongamos que, en dos resultados tales, existiese el mismo número de personas. Si en uno de estos resultados las personas saliesen mucho peor paradas, este sería evidentemente el peor resultado. Sería peor para nadie, pero, como he argumentado, esto no demuestra que no pueda ser peor.

Volvamos a A y B. ¿Cuál resultado sería mejor? Es evidentemente malo que, en B, las personas resulten menos favorecidas. ¿Podría ser esto moralmente compensado por el hecho de que haya más personas vivas?

Supongamos que creemos que, en el Caso del Niño Feliz, la pareja de mi ejemplo no tiene ninguna razón moral para tenerlo. Podemos creer entonces que, si las personas están en peor situación, esto no puede ser moralmente compensado por un incremento en el número de las personas vivas. Los que piensan esto con frecuencia apelan a

> *El Principio Impersonal de la Media*: Si no intervienen otros factores, el mejor resultado es aquel en el que la vida de las personas, por término medio, va lo mejor posible.

Hay economistas que hacen a este principio verdadero por definición [20]. Lo llamo *impersonal* porque no se plantea en términos de *personas afectadas*: no versa sobre lo que sería bueno o malo para las personas a las que nuestros actos afectan. Este principio no supone que, si se causa que existan personas cuya vida valga la pena vivirse, esas personas salen con ello beneficiadas.

La Versión Hedonista de este principio establece que

> Si no intervienen otros factores, el mejor resultado es aquel en el que hay mayor suma neta media de felicidad por vida vivida.

Formulo estas versiones de un modo temporalmente neutral. Hay quienes formulan el Principio de la Media de forma que se

[20] Véase, por ejemplo, Samuelson, p. 551.

refiera sólo a las personas que están vivas después de que hayamos actuado. En esta forma el principio implica absurdamente que sería mejor si, de las personas que están ahora vivas, se matara a todas excepto a las más eufóricas. Según una versión temporalmente neutral del Principio de la Media, si alguien con una vida digna de vivirse muere antes, esto hace que las vidas de las personas, por término medio, vayan peor.

Supongamos a renglón seguido que, en el Caso del Niño Feliz, la pareja de mi ejemplo sí que tiene una razón moral para tenerlo. Creemos que es siempre mejor en sí mismo que se viva una vida más que sea digna de vivirse. Si es esto lo que creemos, sería natural afirmar que, de mis dos resultados, B podría ser mejor que A. La pérdida en calidad de vida podría ser compensada por una ganancia suficiente en el número de vidas vividas. Si suscribimos esta afirmación, tenemos que preguntar, «¿Qué sería una ganancia *suficiente*?».

En caso de ser hedonistas, podríamos formular con facilidad estas preguntas con mayor precisión. Preguntamos

> (1) «Si en uno de dos resultados las personas vivas fuesen menos felices, ¿puede esto ser moralmente compensado por un incremento suficiente en la cantidad de felicidad?».

Si las personas son menos felices, tienen una calidad de vida más baja. Si contestamos Sí a la pregunta (1), tenemos que preguntar

> (2) «¿Cuáles son los valores relativos de la calidad y la cantidad?»

La Versión Hedonista del *Principio Impersonal del Total* nos da una respuesta. Esta afirma

> Si no intervienen otros factores, el mejor resultado es aquel en el que haya la cantidad mayor de felicidad —la suma neta mayor de felicidad menos sufrimiento.

Según este principio, B sería mejor que A, desde el momento en que en B habría una cantidad mayor de felicidad. Aunque las personas B son cada una de ellas menos felices que las personas A, cada una de sus vidas contiene más de la mitad de la felicidad de la que contienen las otras. Como hay el doble de personas B, todas juntas tienen más felicidad que las personas A. (Dos botellas que están llenas hasta más de la mitad contienen más que una botella llena.)

Supongamos, a continuación, que no somos hedonistas. Lo que es para nosotros moralmente importante no es la felicidad sino la calidad de vida. Podemos hacer las mismas preguntas, pero tenemos que utilizar una expresión que no es corriente. Cuando comparamos los valores de la calidad y la cantidad, ¿cuál es la cantidad relevante? Podríamos decir, «la cantidad de vidas vividas que valen la pena vivirse». Pero esto es incorrecto, puesto que ignora la *calidad* de estas vidas de más, o cuánto valen la pena vivirse. La cantidad relevante tiene que ser, como la suma de felicidad, una función tanto del número de estas vidas como de su calidad. Para describir la cantidad relevante, sugiero la expresión «la cantidad de todo lo que hace la vida digna de vivirse».

Reconsideremos A y B. Los hedonistas dirían: «Aunque las personas B son cada una de ellas menos felices que las personas A, juntas tienen más felicidad». Podemos afirmar de manera parecida: «Comparadas con las personas A, cada una de las personas B tiene menos de todo lo que hace la vida digna de vivirse. Pero cada vida en B vale la pena de vivirse más de la mitad de cada vida en A. Como hay el doble de personas B, ellas juntas tienen más de todo lo que hace la vida digna de vivirse».

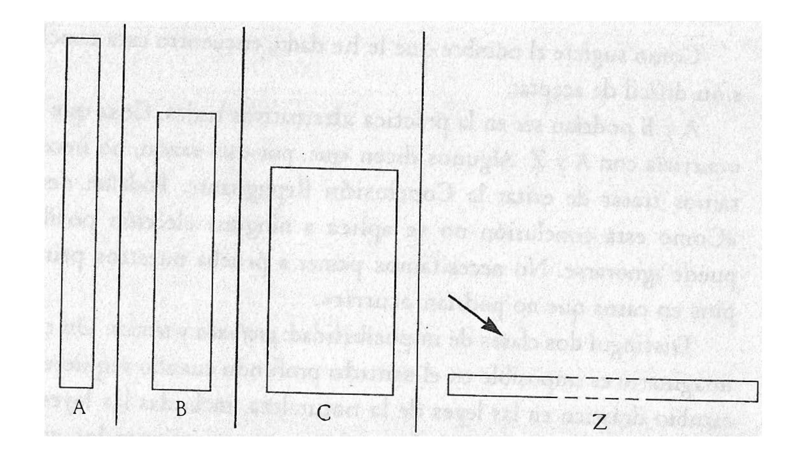
Ahora puedo formular el no hedonista

> *Principio Impersonal del Total*: Si no intervienen otros factores, el mejor resultado es aquel en el que haya la cantidad mayor de todo lo que hace la vida digna de ser vivida.

Si creemos que B sería peor que A, tenemos que rechazar este principio.

131. La conclusión repugnante

Consideremos a continuación el diagrama mayor abajo



Según el Principio Impersonal del Total, igual que B sería mejor que A, C sería mejor que B. Y Z podría ser el mejor de todos. Z es una enorme población cuyos miembros viven vidas que no se elevan mucho sobre el nivel en el que la vida deja de valer la pena. Una vida podría ser así o bien porque tiene suficientes éxtasis como para hacer que parezca que vale la pena soportar sus miserias, o bien porque es de pobre calidad en general. Imaginemos que las vidas en Z son de esta segunda clase más monótona. En cada una de estas vidas hay muy poca felicidad. Pero, si los números son lo suficientemente grandes, este es el resultado con la mayor suma total de felicidad. De forma similar, Z podría ser el resultado en el que haya la mayor cantidad de todo lo que hace que la vida valga la pena. (Podría encontrarse la mayor cantidad de leche en un montón de botellas conteniendo cada una sólo una gota.)

Supongamos a renglón seguido, por una razón que revelaré después, que A tuviera una población de diez mil millones. El Principio Impersonal del Total implica entonces

> *La Conclusión Repugnante*: Para cualquier población posible de al menos diez mil millones de personas, todas con una calidad de vida muy alta,

tiene que haber una población imaginable mucho más grande, cuya existencia, si no intervienen otros factores, sería mejor, aunque sus miembros tengan vidas que apenas sean dignas de vivirse.

Como sugiere el nombre que le he dado, encuentro esta conclusión difícil de aceptar.

A y B podrían ser en la práctica alternativas reales. Cosa que no ocurriría con A y Z. Algunos dicen que, por esta razón, no necesitamos tratar de evitar la Conclusión Repugnante. Podrían decir: «Como esta conclusión no se aplica a ninguna elección posible, puede ignorarse. No necesitamos poner a prueba nuestros principios en casos que no podrían ocurrir».

Distinguí dos clases de imposibilidad: *profunda* y *técnica*. Un caso imaginario es imposible en el sentido profundo cuando requiere un cambio drástico en las leyes de la naturaleza, incluidas las leyes de la naturaleza humana. Hay dos razones para cuestionar los casos que son imposibles en este sentido profundo. Puede que seamos incapaces de imaginar lo que tales casos implicarían. Y algunos dirían que nuestros principios morales precisan ser aceptables solamente en el mundo real [21].

Puede servir de ayuda recordar los imaginarios *monstruos de utilidad* de Nozick. Se trata de personas que obtienen «de todo sacrificio de los demás, ganancias en utilidad enormemente mayores que lo que los demás pierden» [22]. Una persona imaginaria semejante constituye una objeción contra el Utilitarismo de Actos, el cual «parece requerir que todos nosotros nos sacrifiquemos en el estómago del monstruo, para incrementar la utilidad total».

Tal y como Nozick la describe, una persona así es una imposibilidad profunda. La población mundial es ahora de varios miles de millones. Imaginemos la desdicha de todas esas personas si se les niega todo lo que vaya más allá de raciones para no morir de hambre, y todos los demás recursos fueran al monstruo imaginario de

[21] Véase, por ejemplo, la discusión de los diferentes niveles del razonamiento moral en Hare (1) y (2).

[22] Nozick (2), p. 41.

Nozick. Nozick nos pide que supongamos que esta persona imaginaria sería *tan* feliz, o tendría una vida de calidad *tan* alta, que esta es la distribución que produce la mayor suma de felicidad, o la mayor cantidad de todo lo que hace la vida digna de ser vivida. ¿Cómo va a ser esto así, con miles de millones de individuos entregados a la miseria absoluta que podrían ser tan fácilmente aliviados por una pequeña fracción de los vastos recursos de este monstruo? Para que esto sea así, la calidad de vida de este monstruo tiene que ser *millones* de veces más alta que la de cualquiera que conozcamos. ¿Nos lo podemos imaginar? Piensa en la vida de la persona más afortunada que conozcas, y pregúntate cómo tendría que ser una vida para ser un millón de veces más digna de ser vivida. El abismo cualitativo entre semejante vida y las nuestras, en el mejor de los casos, tiene que asemejarse al abismo entre las nuestras, en el mejor de los casos, y la vida de las criaturas que apenas son conscientes —como las «ostras satisfechas» [23] de Platón, si es que ellas *son* conscientes—. Parece una respuesta razonable decir que no podemos imaginar, ni siquiera del modo más vago, la vida de este monstruo de utilidad. Y esto pone en tela de juicio la fuerza del ejemplo. Los utilitaristas del acto podrían decir que, si realmente pudiéramos imaginar cómo sería una vida semejante, no podríamos encontrar convincente la objeción de Nozick. Su «monstruo» parece un ser semejante a Dios. En la imaginaria presencia de tal ser, nuestra creencia en nuestro derecho a la igualdad con él puede empezar a vacilar —del mismo modo que no creemos que los animales inferiores tengan derecho a ser iguales a nosotros.

Esta respuesta tiene alguna fuerza. Pero hasta una imposibilidad profunda puede proporcionar una prueba parcial a nuestros principios morales. No podemos ignorar sin más los casos imaginarios.

Volvamos ahora a mi Z imaginaria. Esta población imaginaria es otro monstruo de utilidad. La diferencia es que la mayor suma de felicidad viene de un vasto incremento no de la calidad de la vida de una persona sino del número de vidas vividas. Y *mi* mons-

[23] Platón (2), 21 c-d.

truo de utilidad ni es imposible en sentido profundo ni es algo que no podamos imaginar. Podemos imaginar lo que sería para la vida de alguien ser apenas digna de ser vivida. Y podemos imaginar lo que sería que hubiera muchas personas con vidas así. Para imaginar Z, simplemente tenemos que imaginar que habría *muchísimas*. Esto lo podemos hacer. Así que el ejemplo no puede cuestionarse como uno que apenas podamos comprender.

En la práctica, no podríamos hacer frente a una elección entre A y Z. Dada una reserva limitada de recursos, no podríamos producir de hecho la mayor suma de felicidad, o la mayor cantidad de todo lo que hace la vida digna de vivirse, produciendo una enorme población cuyas vidas apenas sean dignas de vivirse [24]. Pero esto sería meramente imposible en el sentido técnico. Para suponerlo posible, sólo necesitamos añadir algunos detalles sobre la naturaleza y la disponibilidad de los recursos. Como sería meramente imposible en el sentido técnico afrontar una elección entre A y Z, esto no debilita la comparación como una prueba para nuestros principios. Las Elecciones de Diferente Número hacen surgir la pregunta de si la pérdida en la calidad de vida podría ser *siempre* moralmente compensada por una ganancia suficiente en la cantidad o de felicidad o de cualquier cosa que haga la vida digna de vivirse. Esta es la pregunta que la comparación de A y Z plantea del modo más claro. Si estamos convencidos de que Z es peor que A, tenemos poderosas razones para resistirnos a unos principios que implican que Z es mejor. Tenemos poderosas razones para resistirnos al Principio Impersonal del Total.

[24] Según algunas versiones de la *Ley de la Utilidad Marginal Decreciente*, esto es justo lo que está implicado. Según estas versiones, cada unidad de recursos produce más utilidad si se le da a la gente que está en peores condiciones, de manera que la distribución más productiva será aquella en que la vida de cada uno sea apenas digna de vivirse. Hay aquí un descuido obvio. Se necesitan grandes cantidades de recursos para hacer que la vida de cada persona alcance siquiera el nivel donde la vida empieza a ser digna de vivirse. Tales recursos no ayudan a producir la mayor suma neta causalmente posible de utilidad cuando se usan meramente para impedir que la gente *extra* tenga una vida *digna de terminar* (o tenga falta de utilidad neta).

Alguien podría decir: «No es así. Este principio incluye la frase *si no intervienen otros factores*. Pero siempre habría algún otro factor. Podemos por consiguiente ignorar la Conclusión Repugnante».

Esto no es plausible. ¿Qué otro principio moral tiene que ser infringido por el ocurrir de Z? Podría afirmarse que este infringiría algún principio sobre la justicia entre generaciones. Pero esto es irrelevante para nuestra pregunta en su forma más pura. Estamos preguntando si, en caso de que ocurriera Z, esto sería mejor que si ocurriera A. Podríamos imaginar una historia en la que sólo ocurrieran resultados semejantes a Z. Las personas en Z no serían entonces menos favorecidas que nadie que alguna vez viviera. Si creemos que Z sería peor que A, esto no podría ser aquí porque la ocurrencia de Z conllevara injusticia.

Hay otra cuestión más importante. Reconsideremos el Problema de la No-Identidad. Hay quienes sugieren que podemos resolver este problema con una apelación a los derechos de las personas. Pero, como muestra el caso de la Reducción, no es así. Si eliminamos con la imaginación el Problema de la No-Identidad, la objeción a nuestra elección de la Reducción Menor apelaría a nuestro Principio de Benevolencia. Para resolver el Problema de la No-Identidad, tenemos que revisar este principio. Tenemos que encontrar lo que llamo Teoría X.

Ocurre lo mismo si queremos evitar la Conclusión Repugnante. No deberíamos tratar de evitar esta conclusión con una apelación a principios que cubren una parte diferente de la moralidad. Esta conclusión es *intrínsecamente* repugnante. Y esta conclusión está implicada por el Principio Impersonal del Total, que es una versión especial del Principio de Beneficencia. Para evitar la Conclusión Repugnante, tenemos que tratar de hacer patente que deberíamos rechazar esta versión. Tenemos que tratar de encontrar una versión mejor: la Teoría X.

18

LA CONCLUSIÓN ABSURDA

Necesitamos una teoría que resuelva el Problema de la No-Identidad, y a la vez evite la Conclusión Repugnante. Como veremos, varias teorías consiguen uno de estos objetivos a costa de su fracaso en lograr el otro.

132. Una supuesta asimetría

Hay otro objetivo que, de acuerdo con muchos, deberíamos tratar de conseguir. Consideremos

> *El Niño Desgraciado.* Una mujer sabe que, si tiene un hijo, va a tener tantas enfermedades que su vida será peor que nada. Nunca se desarrollará, sólo vivirá unos cuantos años, y sufrirá un dolor que no va a poder ser aliviado completamente.

Aunque rechacemos la expresión «peor que nada», está claro que estaría mal concebir a sabiendas un hijo así. Y lo incorrecto no provendría mayormente de los efectos sobre otras personas. Lo incorrecto provendría sobre todo de que la calidad de vida de este niño sería previsiblemente atroz.

Recordemos a continuación el Caso del Niño Feliz. La pareja de mi ejemplo puede suponer que, si tienen este niño, no será peor ni para ellos ni para otras personas. ¿Cuál es la relación entre estos dos casos?

Algunos autores afirman que, mientras que tendríamos el deber de no concebir al Niño Desgraciado, la pareja de mi ejemplo no tiene el de concebir al Niño Feliz. Sería simplemente mejor desde el punto de vista moral que tuvieran ese hijo.

Y muchos negarían incluso esta última afirmación. Estas personas creen que, mientras que sería incorrecto tener al niño desgraciado, la pareja de mi ejemplo *no* tiene ninguna razón moral para tener al niño feliz [25]. Se ha denominado a este modo de pensar *la Asimetría* [26].

Si aceptamos esta opinión, entonces tendremos un tercer objetivo. Además de resolver el Problema de la No-Identidad y de evitar la Conclusión Repugnante, también tenemos que explicar la Asimetría. ¿Qué teoría conseguiría estos objetivos?

133. Por qué el método contractual ideal no proporciona ninguna solución

¿Podemos recurrir al método de razonamiento moral que denominaré *Contractualismo Ideal*? Esto lo defienden muchos autores, siendo más célebre de todos Rawls. Según esta concepción, los mejores principios morales son los que sería racional que eligiéramos como principios a seguir en nuestra sociedad. Para asegurar la imparcialidad, preguntamos qué principios deberíamos elegir si no conociéramos ciertos hechos especiales sobre nosotros mismos.

Algunos afirman que si aplicamos este método a la cuestión de cuántas personas debería haber, podremos explicar por qué deberíamos rechazar tanto el Principio Impersonal del Total como la Conclusión Repugnante. Podremos explicar por qué B sería peor que A, y por qué, de todos los resultados, Z sería el peor. Aunque no conociéramos ciertos hechos especiales sobre nosotros mismos, elegiríamos un principio que produjese A antes que B. Entonces estaríamos seguros de resultar más favorecidos.

[25] Véase Narveson (1) y (2), y J. Bennett y M. Warren en Sikora y Barry. Hay otros muchos ejemplos.

[26] En McMahan (1). En toda la Cuarta Parte me hallo muy en deuda con el trabajo inédito de J. McMahan, y con muchas discusiones de estas cuestiones.

Como defienden varios autores [27], esta no es una forma aceptable de defender un principio acerca del número de personas que debería haber. Este método de razonamiento moral recurre a lo que sería racional que eligiéramos, en términos del propio interés. Es esencial a este método que no sepamos si aguantaríamos lo más duro del principio elegido. Así, si el principio elegido perjudicase a las mujeres, tendríamos que imaginar que no conocemos nuestro sexo. Sólo este *velo de ignorancia* hace a nuestra elección imparcial en la forma que la moralidad requiere.

A tenor del razonamiento sugerido arriba, sería racional elegir un principio que produjese A antes que B, puesto que entonces, con seguridad, saldríamos más favorecidos. Este razonamiento da por sentado que, sea cual sea el principio que se siga, nosotros sin duda alguna existiremos. Este supuesto viola el requisito de imparcialidad. El principio que elegimos afecta a cuántas personas existen. Si damos por hecho que nosotros sin duda alguna existiremos sea cual sea el principio que elijamos, esto es como dar por hecho, cuando elegimos un principio que perjudicaría a las mujeres, que nosotros seremos sin duda alguna varones.

Hay autores que sugieren que deberíamos cambiar nuestro supuesto. Cuando preguntamos qué principio sería racional elegir, deberíamos imaginar que no sabemos si alguna vez existiremos. Si elegimos un principio que producirá una población más pequeña, esto incrementará las probabilidades de que nosotros nunca existamos. Se ha dicho que con este cambio en nuestro supuesto sería racional elegir el Principio Impersonal del Total.

Pero no podemos cambiar nuestro supuesto de la manera apuntada. Podemos imaginar una historia posible diferente, en la que

[27] Véase, por ejemplo, Barry (3).

nosotros nunca existimos. Pero no podemos suponer que, en la historia real del mundo, podría ser verdadero que nosotros nunca existamos. Por eso no podemos preguntar qué sería racional elegir según este supuesto.

El método contractual ideal no puede aplicarse a determinadas partes de la moralidad. Así, como afirma Rawls, no puede aplicarse de forma verosímil a la cuestión de cómo debemos tratar a otros animales [28]. Considero que, por las diferentes razones que acabo de dar, este método no puede aplicarse de forma verosímil a la elección del principio relativo a cuántas personas debería haber. Cuando estamos discutiendo esta cuestión, este método no es imparcial, a no ser que imaginemos algo que bajo ningún concepto podemos imaginar.

Un defensor de este método podría rechazar mi afirmación sobre la imparcialidad. Podría seguir creyendo que, al aplicarlo, podemos dar por supuesto que sin duda alguna existiremos. Consideremos dos posibilidades para la última generación de la historia humana. (Considerar la última generación simplifica el caso.) La historia podría acontecer en dos formas, antes de que súbitamente terminase a causa de la explosión del sol:

> En *Infierno Uno*, la última generación consta de diez personas inocentes, cada una de las cuales sufre un gran dolor durante cincuenta años. Las vidas de estas personas son mucho peores que nada. Si pudieran se darían muerte a sí mismas.
>
> En *Infierno Dos*, la última generación consta no de diez sino de diez millones de personas inocentes, cada una de las cuales sufre un gran dolor también durante cincuenta años, menos un día.

Si suponemos que nosotros existiremos sin duda alguna en uno de estos dos infiernos, evidentemente sería racional, en términos del propio interés, preferir el Infierno Dos, puesto que entonces sufriríamos un día menos. ¿Deberíamos concluir que el Infierno Dos sería mejor, en el sentido que tiene relevancia moral?

[28] Rawls, p. 17, y pp. 504-12.

¿Actuaría Satanás menos mal si este fuese el Infierno que pusiera en obra?

La respuesta a las dos preguntas es No. El Infierno Dos es mejor en un aspecto. La cantidad de dolor por persona sería muy ligeramente menor; se reduciría menos del 0,01 por ciento. Pero este hecho queda moralmente compensado por el enorme incremento del número de personas que sufren dolor, y soportan vidas que son mucho peores que nada. En el Infierno Dos la cantidad de sufrimiento es casi un millón de veces mayor.

El ejemplo muestra que, según el método de razonamiento moral propuesto, somos llevados a ignorar completamente lo que tiene que admitirse que tiene como mínimo alguna significación moral. Somos llevados a ignorar el hecho de que, en uno de estos resultados, habría muchísimo más sufrimiento. Semejante método de razonamiento moral no puede ser aceptable. Tenemos que encontrar algún otro modo de lograr nuestros objetivos.

134. El principio estrecho de las personas afectadas

Si es preciso que evitemos la Conclusión Repugnante, tendremos que rechazar el Principio Impersonal del Total. Tendremos que afirmar que este principio describe mal la parte de la moralidad que tiene que ver con la beneficencia, o el bienestar humano.

Podríamos decir que este principio adopta la forma equivocada, puesto que trata a las personas como meros *contenedores* o *productores* de valor. Aquí tenemos un fragmento en que este rasgo queda especialmente en evidencia. En su libro *The Technology of Happiness* [*La tecnología de la felicidad*], Mackaye escribe:

> «Igual que se requiere una caldera para utilizar la energía potencial del carbón en la producción de vapor, se requieren seres sensibles para convertir el potencial de felicidad residente en un área de tierra dada en felicidad efectiva. Y de la misma manera que el Maquinista elegirá calderas que sean máximamente eficientes en la conversión del vapor en energía, la Justicia elegirá seres sensibles



que tengan la máxima eficiencia a la hora de convertir recursos en felicidad» [30].

Podría decirse que este *Modelo de la Producción de Vapor* es una distorsión grotesca de esta parte de la moralidad. Podríamos recurrir a

La Restricción de las Personas Afectadas: Esta parte de la moralidad, la parte que tiene que ver con el bienestar humano, debería ser explicada enteramente en términos de lo que sería bueno o malo para las personas a las que nuestros actos afectan.

Esta es la tesis propuesta por Narveson [31]. Según su concepción, no es bueno que las personas existan por el hecho de que sus vidas contengan felicidad. Antes bien, la felicidad es buena por el hecho de que es buena para las personas.

También afirma Narveson que, al causar que alguien exista, no podemos estar beneficiando a esta persona. Por consiguiente, afirma que de los dos modos de incrementar la suma de felicidad —hacer a la gente feliz, y hacer gente feliz— sólo el primero es bueno para las personas. Como esta parte de la moralidad sólo tiene que ver con lo que es bueno o malo para las personas, el segundo modo de incrementar la felicidad es, afirma él, moralmente neutral.

Como sostuve en el capítulo 16, tenemos que rechazar la idea de que lo que es malo tiene que ser malo para alguien. Puede parecer que, al rechazar esta idea, estamos rechazando la Restricción de las Personas Afectadas. Pero no es así. A diferencia de Narveson, nosotros podemos creer que, al causar que exista alguien que va a tener una vida digna de vivirse, estamos con ello beneficiando a esta persona. El Apéndice G defiende esta creencia. Si la aceptamos, podemos explicar, en términos de personas afectadas, por qué tenemos una razón moral para no producir ciertos efectos, aunque no

[30] Mackaye. No tengo la referencia de la página. Y como cito de memoria la segunda frase, puede que sea inexacta. Este libro olvidado divertiría e instruiría a los utilitaristas, y encantaría a sus oponentes.

[31] En Narveson (1). Buena parte del reciente debate debe su existencia a este trabajo pionero. Véase también Narveson (2).

vayan a ser malos para nadie. Podemos explicar por qué tenemos una razón moral para no elegir la Política Arriesgada, o la Reducción.

Si creemos que causar que se exista puede beneficiar, tenemos que decidir entre al menos tres Principios de Beneficencia diferentes. A diferencia del Principio de Utilidad, estos tres principios no pretenden cubrir el todo de la moralidad. Como incluyen la expresión «si no intervienen otros factores», nos permiten recurrir a otros principios, como el de Igualdad. Podemos afirmar también que no tenemos la obligación de beneficiar a los demás cuando esto exija de nosotros un sacrificio demasiado grande, o una interferencia demasiado grande en nuestras vidas.

Estos tres principios afirman todos:

(1) Si se causa que exista alguien, y tiene una vida que vale la pena vivir, esta persona es con ello beneficiada. Este beneficio es mayor si su vida es más digna de vivirse.

(2) Si no intervienen otros factores, es incorrecto hacer a sabiendas una elección que haga el resultado peor.

(3) Si no intervienen otros factores, uno de dos resultados sería peor si fuese peor para las personas.

Estos principios difieren en sus afirmaciones sobre lo que hace a un resultado peor para las personas. Supongamos que estamos comparando los resultados *X* e *Y*. Llamemos a las personas que existirán en el resultado X *las personas-X*. De estos dos resultados, llamemos a X

«peor para las personas» en el sentido *estrecho* si la ocurrencia de X más bien que la de Y sería o peor, o mala para las personas-X.

De acuerdo con la afirmación (1), al causarse que exista, alguien puede ser beneficiado. Como escribo en el Apéndice G, no necesitamos afirmar que este resultado sea *mejor* para esta persona que la alternativa. Esto implicaría la afirmación inverosímil de que, si esta persona no hubiera existido nunca, esto habría sido peor para

ella. Podemos afirmar en vez de eso que, si se hace existir a alguien, esto puede ser *bueno* para él. Podemos afirmar de una manera similar que, si al Niño Desgraciado se le hace existir, esto es malo para él. Esto es porque, en la definición que se acaba de dar, incluyo la expresión «o mala para».

Si usamos «peor para las personas» en el sentido recién definido, las afirmaciones que van de (1) a (3) conforman el Principio *Estrecho* de Beneficencia de las Personas Afectadas, o —para abreviar— el *Principio Estrecho*.

135 Por qué no podemos apelar a este principio

El Principio Estrecho es intuitivamente plausible. Es natural asumir que, si una elección no será mala para nadie que alguna vez viva, nuestro Principio de Beneficencia no debería condenarla. Pero, si apelamos a la Restricción de las Personas Afectadas, tenemos que rechazar el Principio Estrecho. Este principio no puede resolver el Problema de la No-Identidad. Necesitamos explicar por qué tenemos una razón moral para no elegir la Política Arriesgada, o la Reducción. De acuerdo con el Principio Estrecho, no tenemos tal razón, puesto que sabemos que estas elecciones no serán peores para nadie.

Hay otras objeciones al Principio Estrecho. Una es que, si este es nuestro Principio de Beneficencia, tenemos que aceptar parte de la Conclusión Repugnante. Si lo que sucede es Z en vez de A, esto no sería o peor o malo para cualquiera de las personas que alguna vez vivan. Si apelamos tanto a la Restricción de las Personas Afectadas como al Principio Estrecho, tenemos que afirmar que Z no sería peor que A. La mayoría de nosotros encontraría esto difícil de creer.

Otra objeción es que el Principio Estrecho puede implicar contradicciones. Consideremos estos dos resultados:

(1) A diferencia de otras personas existentes, Jack lleva una vida que es mucho peor que nada. Jill nunca existe.
(2) A diferencia de otras personas existentes, Jill lleva una vida que es mucho peor que nada. Jack nunca existe.

Si lo que ocurre es (1) en vez de (2), esto será malo para Jack, y no será bueno para Jill. (1) sería, por tanto, peor que (2) para las personas-(1). Por un razonamiento similar, (2) sería peor que (1) para las personas-(2). El Principio Estrecho implica la conclusión contradictoria de que cada uno de estos resultados sería peor que el otro.

Resulta tentador combinar la Restricción de las Personas Afectadas y el Principio Estrecho. Pero, si apelamos a esta concepción, no podremos resolver el Problema de la No-Identidad, tendremos que aceptar parte de la Conclusión Repugnante, y seremos llevados a contradicciones. Hemos aprendido, por tanto, que no podemos apelar a esta concepción. Esto es progreso, de una clase negativa.

Hay otra manera en que nuestro progreso ha sido negativo. Muchos aceptan la Asimetría. Estas personas quieren explicar por qué, aunque sería incorrecto concebir a sabiendas al Niño Desgraciado, la pareja de mi ejemplo no tiene ninguna razón moral para concebir al Niño Feliz. Como no podemos recurrir a la opinión que acabamos de mencionar, argumento sólo en una nota que esta opinión *habría* aportado la mejor explicación de la Asimetría [32].

[32] Según la supuesta Asimetría, sería incorrecto tener el Niño Desgraciado, pero no hay ninguna razón moral para tener el Niño Feliz. ¿Cómo puede explicarse esto?

Algunos autores afirman que hacer que alguien exista no puede ser ni bueno ni malo para él. Si esto es así, se explicaría por qué no hay ninguna razón moral para tener al Niño Feliz. Pero implicaría que, al tener el Niño Desgraciado no podemos estar haciendo nada que sea malo para él. ¿Cuál es la objeción? Si tener este hijo no puede ser malo para él, la mayor parte de la incorrección de este caso tiene que venir de su sufrimiento. Tenemos que apelar a

> *El Principio del Sufrimiento Total*: Si no intervienen otros factores, no debemos incrementar la suma de sufrimiento.

Puede ser difícil aceptar este principio pero rechazar completamente el Principio de la Felicidad Total. Pero si aceptamos ese principio nuestra pareja tiene una razón moral para tener el Niño Feliz. Aunque tenerlo no puede ser bueno para él, incrementaría la suma de felicidad.

Puede objetarse que el sufrimiento y la felicidad son moralmente diferentes —que las razones morales que tenemos para impedir el primero pesan mucho más

136. Los dos Principios Amplios de las Personas Afectadas

Describiré ahora dos Principios de Beneficencia de las Personas Afectadas más. Como el Principio Estrecho, estos dos afirman:

que las que tenemos para promover la segunda. Pero esto no puede explicar la Asimetría a menos que no tengamos *ninguna* razón moral para promover la felicidad. Y si hemos aceptado el Principio del Sufrimiento Total, no parece plausible rechazar completamente su análogo para la felicidad. [Véase Griffin (4).]

Hay una manera mejor de explicar la Asimetría. Podríamos (1) apelar a la Restricción de las Personas Afectadas, (2) afirmar que hacer que alguien exista puede ser o bueno o malo para él, y (3) apelar al Principio Estrecho. De acuerdo con el Principio Estrecho, es incorrecto, si no intervienen otros factores, hacer lo que sería o malo para, o peor para, la gente que alguna vez viva. Es, por tanto, incorrecto tener el Niño Desgraciado, puesto que esto sería malo para él. Pero de ningún modo es incorrecto dejar de tener el Niño Feliz. Es cierto que si mi pareja tiene este hijo va a ser mejor para él. Pero, si no lo tiene, no será malo para él. Y, en el caso descrito, no será malo para nadie más. Por eso no hay ninguna razón moral para tenerlo.

Esta me parece la mejor explicación de la Asimetría. Deberíamos notar que el Principio Estrecho no conlleva la afirmación familiar de que nuestra obligación de no hacer daño es más fuerte que nuestra obligación de beneficiar a los demás. Tal vez deseemos añadir esta afirmación a nuestro Principio Estrecho de Beneficencia —añadir, como hizo Ross, algún principio más fuerte sobre la Maleficencia. Pero el Principio Estrecho no hace semejante distinción—. Si va a haber en el futuro alguien a quien hayamos dejado de dar un beneficio, nuestro no hacerlo así será peor para esta persona. Si no intervienen otros factores, hemos actuado incorrectamente según el Principio Estrecho.

La distinción entre los Principios Amplio y Estrecho es una distinción nueva, abierta por la creencia de que, al hacer que alguien exista, podemos estar haciendo algo que es bueno o malo para él. Esta creencia rompe la implicación corriente de que, si un suceso fuese bueno para la gente, la no ocurrencia del mismo sería peor para la gente. Con esta implicación rota, los Principios Amplio y Estrecho divergen.

Deberíamos tomar nota, por último, de que podemos justificar nuestra apelación al Principio Estrecho. Según este, tiene significación moral que, al hacer que alguien exista, lo que hacemos sea bueno para esta persona. J. Sterba ha objetado que el Principio Estrecho no explica aquí la Asimetría, puesto que se limita a reformularla. Se puede responder a esta objeción. Podríamos justificar nuestra apelación al Principio Estrecho apelando a varias teorías sobre la naturaleza de la moralidad, o del razonamiento moral. Un ejemplo es la teoría de Scanlon. Como he dicho, esta afirma que la fuente de la motivación moral es «el deseo de poder justificar nuestras propias acciones ante los demás recurriendo a razones que ellos no podrían

(1) Si se causa que exista alguien, y tiene una vida que vale la pena vivir, esta persona es con ello beneficiada. Este beneficio es mayor si la vida de esta persona es más digna de vivirse.
(2) Si no intervienen otros factores, es incorrecto hacer a sabiendas una elección que produce el resultado peor.
(3) Si no intervienen otros factores, uno de dos resultados sería peor si fuese peor para las personas.

Llamemos al resultado X

«peor para las personas» en el sentido *amplio* si la ocurrencia de X sería menos buena para las personas-X que la ocurrencia de Y para las personas-Y.

En Elecciones de Diferente Número «menos bueno para» es ambiguo. Llamemos a X

«peor para las personas» en el sentido *amplio del total* si el beneficio neto total dado a las personas-X por la ocurrencia de X sería menor que el beneficio neto total dado a las personas-Y por la ocurrencia de Y.

Llamemos a X

«peor para las personas» en el sentido *amplio de la media* si el beneficio neto medio por persona dado a las personas-X por la ocurren-

rechazar de forma razonable» [Scanlon (3), p. 116]. Según esta teoría, un acto es incorrecto sólo si afecta a alguien de un modo que le da una queja que no se puede contestar. En la nota 17 menciono otras varias teorías morales que podrían justificar la apelación al Principio Estrecho. Según ellas, es moralmente significativo que, si la pareja de mi caso deja de tener el Niño Feliz, no habrá ningún querellante.

Aunque aporte la mejor explicación de la Asimetría, hemos visto que tenemos que rechazar la concepción que combina el Principio Estrecho y la Restricción de las Personas Afectadas. Tenemos que rechazar la idea de que, para que un acto quede expuesto a una objeción moral, tiene que haber algún querellante. Si elegimos la Reducción, esto causará más tarde un gran declive en la calidad de vida. Pero los que vivan antes de este declive deberán su existencia a nuestra elección. Como no lamentarán su existencia, no lamentarán nuestra elección. No habrá querellantes. Pero *hay* una objeción moral a nuestra elección.

cia de X sería menor que el beneficio neto medio por persona dado a las personas-Y por la ocurrencia de Y.

Si empleamos «peor para las personas» en este sentido amplio del total, las afirmaciones de la (1) a la (3) conforman el Principio *Amplio del Total* de las Personas Afectadas. Si empleamos «peor para las personas» en su sentido amplio de la media, de la (1) a la (3) conforman el Principio *Amplio de la Media* de las Personas Afectadas. (Como explico en la nota [33], si rechazamos la tesis (1), negando que causar que se exista pueda beneficiar, estos dos Principios Amplios coinciden con el Principio Estrecho.)

Según el Principio Amplio del Total, el mejor resultado es el que da a la gente la mayor suma neta total de beneficios —la mayor suma de beneficios menos cargas—. Según el Principio Amplio de la Media, el mejor resultado es el que da a la gente la mayor suma neta media de beneficios por persona. Al apelar a estos principios, podemos ampliar el uso de «beneficio» del modo defendido en la Sección 25. El acto que beneficia más a la gente es el acto cuya consecuencia es que la gente recibe la mayor suma neta total o media de beneficios. Es irrelevante que muchos otros actos también vayan a ser partes de la causa de la recepción de estos beneficios.

Los dos Principios Amplios resuelven el Problema de la No-Identidad. Los dos implican (C), la Tesis de la Calidad del Mismo Número. Y la explican de una manera más familiar. ¿Cuál es nuestra razón moral para no elegir la Reducción? Según los Principios Amplios, esta elección beneficia a los que vivan más adelante, puesto que sus vidas valen la pena, y ellos deben su existencia a nuestra elección. Pero si hubiéramos elegido la Conservación otras personas habrían vivido más adelante, y habrían tenido una calidad de vida más alta. Según los Principios Amplios, si se causa que existan unas

[33] Si hacer que alguien exista no puede ser ni bueno ni malo para él, las únicas personas cuyos intereses cuentan, cuando estamos comparando dos resultados, serán las que existan en los dos. Todas las elecciones entre pares de resultados son, para propósitos morales, como las Elecciones del Mismo Número, en que coinciden los Principios Estrecho y Amplio.

personas, y tienen una calidad de vida más alta, estas personas son más beneficiadas. La objeción a la Reducción es que, aunque beneficie a los que vivan más adelante, les beneficia menos de lo que la Conservación habría beneficiado a los que hubieran vivido más adelante. ¿Por qué se debería haber esperado la Muchacha de 14 Años y haber tenido su hijo más tarde? Porque tener un hijo ahora probablemente beneficiaría a este niño menos de lo que tener un hijo más tarde beneficiaría a ese otro niño. Una afirmación parecida se aplica a nuestra elección de la Política Arriesgada. Esta elección beneficia a la gente que más tarde muere en la catástrofe, pero la beneficia menos de lo que la Política Segura habría beneficiado a las otras personas que hubieran vivido más tarde.

Estas afirmaciones no conllevan esa especie de truco verbal que antes deseché. Cuando sometemos a discusión a la Muchacha de 14 Años, las palabras «su hijo» y «él» pueden usarse para incluir todos los diferentes hijos que esta muchacha podría tener. Esto aporta un sentido en el que es verdadero que, si esta muchacha tiene su hijo ahora, esto será peor para él. Pero, en el sentido en el que esta afirmación es verdadera, no trata sobre lo que es bueno o malo para personas concretas. Esta afirmación tiene, por tanto, que apelar a un nuevo principio, que necesita explicarse y justificarse.

Esto no es verdadero de los dos Principios Amplios. Como hemos visto hace un momento, cuando decimos que uno de dos resultados es «peor para las personas» en los dos sentidos amplios, nuestra afirmación *versa* sobre lo que sería bueno o malo para personas concretas. Los Principios Amplios no son del todo familiares. No pueden pretender ser nuestro principio corriente sobre los efectos en los intereses de las personas —lo que denomino nuestro Principio de Beneficencia—. Esto es porque ahora estamos suponiendo que, al causar que alguien exista, con ello podemos beneficiar a esa persona. Nuestro principio corriente no nos dice cuál es la importancia moral de tales beneficios. Necesitamos ampliar nuestro principio corriente para que conteste esta pregunta. Y hay al menos tres respuestas que vale la pena considerar. Son las que nos dan los dos Principios Amplios y el Principio Estrecho. Podemos por consiguiente afirmar que estos no son principios de una nueva

clase, que necesiten explicarse y justificarse. Exponen tres maneras en que podemos verosímilmente ampliar nuestro principio corriente, como necesitamos hacer si suponemos que causar que se exista puede beneficiar.

Como los Principios Amplios son ampliaciones de nuestro principio corriente, resuelven el Problema de la No-Identidad en las Elecciones del Mismo Número. Explican la tesis C de una forma satisfactoria. Pero esto no demuestra que ninguno de ellos sea aceptable. Tenemos que inquirir lo que implican en las Elecciones de Diferente Número.

Volvamos a los resultados diferentes A y B. Comparado con A, hay en B el doble de personas, que resultan menos favorecidas. Pero las vidas de las personas-B son dignas de vivirse en una medida de más de la mitad de lo que lo son las vidas de las personas-A. Si ocurriera B en vez de A, esto sería «mejor para las personas» en el sentido amplio del total. B sería «menos bueno» para *cada una* de las personas-B de lo que lo sería A para cada una de las personas-A. Pero como cada persona-B se beneficiaría en una medida de más de la mitad de cada persona-A, y habría el doble de personas-B, ellas juntas se beneficiarían más, o recibirían un beneficio total mayor. (En el mismo sentido, si diésemos a cada una de dos personas cuatro años más de vida, juntas recibirían un beneficio mayor que el recibido por una tercera persona a quien le diéramos cinco años más de vida.) El Principio Amplio del Total implica de este modo que, si no intervinieran otro factores, B sería mejor que A.

Por el mismo razonamiento, C sería mejor que B. Y Z podría ser el mejor. Si Z ocurriera, cada una de las personas-Z se beneficiaría con ello muy poco. Pero, si Z fuese suficientemente grande, las personas-Z recibirían juntas el beneficio total mayor. El principio Amplio del Total implica de este modo la Conclusión Repugnante. Como queremos evitarla, tenemos que afirmar que deberíamos rechazar este principio.

(Como los he expuesto, los dos Principios Amplios de las Personas Afectadas amplían nuestro uso corriente de la palabra

«beneficio». Podría objetarse que, al ampliar este uso de «beneficio», mis principios no satisfacen enteramente la Restricción de las Personas Afectadas. Podríamos recurrir a diferentes versiones, que se hallan más cerca de nuestro uso corriente de «beneficio».) Podríamos afirmar

> Si hacemos X en vez de Y, esto beneficia más a las personas en el sentido *amplio del total* si el que hagamos X en vez de Y les da a las personas-X un beneficio total neto más grande que el que hagamos Y en vez de X les habría dado a las personas-Y,

y

> Si hacemos X en vez de Y, esto beneficia más a las personas en el sentido *amplio de la media* si el que hagamos X en vez de Y les da a las personas-X un beneficio medio por persona mayor que el que hagamos Y en vez de X les habría dado a las personas-Y.

Podríamos entonces afirmar que, si no intervienen otros factores, uno de dos resultados sería mejor cuando provocar este resultado beneficiase más a las personas. Estas dos definiciones nos dan diferentes versiones de los Principios Amplio del Total y Amplio de la Media de las Personas Afectadas.

Esta versión del Principio Amplio del Total también implica la Conclusión Repugnante. Si hubiera bastantes personas-Z, provocar Z en vez de A les daría a las personas-Z un beneficio neto total mayor que el que provocar A en vez de Z les daría a las personas-A. Sorprendentemente, esta versión del Principio Amplio *de la Media* puede implicar también la Conclusión Repugnante. Supongamos que hemos provocado A. Ahora podríamos hacer frente a una nueva elección. Supongamos que, en poco tiempo, podríamos cambiar A en B. Podría ser «mejor para las personas» en el nuevo sentido amplio de la media que hiciésemos ese cambio. Este cambio añadiría a la población existente el mismo número de gente. La mitad previamente existente sufriría un declive de su calidad de vida. Este cambio sería peor para ellos. Pero les traería un beneficio mayor a la mitad que empieza a existir. Cambiar de A a B les daría, por tanto,

a las personas-B un beneficio neto medio por persona. Y podría darles un beneficio medio por persona mayor que el que no hacer este cambio daría a las personas-A.

Para ilustrar este punto, fingiré que hay exactitud. Supongamos que el nivel en A es 100, y el nivel en B es 76. En el cambio de A a B cada persona de la mitad previamente existente pierde 24, y cada persona de la mitad que empieza a existir gana 76. El cambio de A a B, por consiguiente, les da a las personas-B un beneficio neto medio por persona de (76-24)/2, o 26. No hacer el cambio le daría a cada una de las personas-A un beneficio de 24. Cambiar de A a B les daría por consiguiente a las personas-B un beneficio medio por persona mayor que el que no hacer este cambio daría a las personas-A. Según la definición dada arriba, el Principio Amplio de la Media implica que B sería mejor que A. Por el mismo razonamiento, C sería mejor que B, D mejor que C, E mejor que D, y así sucesivamente. Estas afirmaciones y otras similares, implican indirectamente que Z sería mejor que A.

Otra objeción a estos principios es que a menudo implican contradicciones. Esto ocurre con el Principio Amplio de la Media en el caso que acabamos de discutir. Supongamos que, comparado con la población de B en el nivel 76, C tendría el doble de gente en el nivel 58. Según el Principio Amplio de la Media, como he mostrado, B sería mejor que A. Por un razonamiento similar, C sería mejor que B. Esto implica, indirectamente, que C sería mejor que A. Pero el Principio Amplio de la Media implica directamente que A sería mejor que C. De acuerdo con este principio, A sería mejor y peor que C. Pruebo este extremo en la nota [34].

[34] En el cambio de B a C, la mitad previamente existente pierde por persona 18, y la mitad nuevamente existente gana por persona 58. Por consiguiente, el cambio de B a C da a las personas C un beneficio medio neto por persona de (58-18) : 2, o 20. Dejar de hacer este cambio daría a cada una de las personas B un beneficio de 18. Cambiar de B a C daría, por tanto, a las personas C un beneficio medio mayor por persona que el que dejar de hacer este cambio daría a las personas B. Según el Principio Amplio de la Media, como se formula arriba, C sería mejor que B. (Al calcular el beneficio medio en el cambio de B a C, yo divido por 2 puesto que los dos grupos son igual de grandes. Este método taquigráfico rinde

Podríamos revisar estos principios para evitar tales contradicciones. Pero, como explica esa nota, los principios serían entonces inadecuados. Puesto que esto es así, y *ambos* principios implican la Conclusión Repugnante, estas dos versiones de los Principios Amplios son inferiores a las versiones expuestas anteriormente. (En lo que sigue me refiero a estas versiones anteriores.)

De nuevo hemos hecho un progreso, uno de tipo negativo. El Principio Impersonal del Total implica la Conclusión Repugnante. En su versión hedonista, este principio afirma que, si no intervienen otros factores, el mejor resultado es aquel en el que habría la mayor suma neta de felicidad. La otra versión de este principio sustituye «felicidad» por «todo lo que hace la vida digna de vivirse».

Como queremos evitar la Conclusión Repugnante, tenemos que decir que deberíamos rechazar el Principio Impersonal del Total. Narveson sugirió una razón para rechazarlo. Podemos recurrir a la Restricción de las Personas Afectadas: la afirmación de que cualquier principio de beneficencia tiene que adoptar una forma de personas afectadas.

la misma conclusión que el cálculo que este principio explícitamente exige. Este añadiría primero la suma neta total de beneficios menos pérdidas a todos estos miles de millones de personas, y luego dividiría esa suma por el número total de personas. Observaciones similares se aplican a mis otros cálculos taquigráficos.) Como este principio implica que C sería mejor que B, que sería mejor que A, implica indirectamente que C sería mejor que A. Pero en un cambio de A a C el cuarto previamente existente perdería 42, y los tres cuartos nuevamente existentes ganarían 58. Por tanto, el cambio de A a C les da a las personas C un beneficio medio neto de (58 x 3 – 42) : 4, o 33. Cambiar de A a C daría por consiguiente a las personas C un beneficio medio por persona más pequeño que el que les daría a las personas A el dejar de hacer este cambio. De acuerdo con este principio, aunque C es indirectamente mejor que A, C es peor que A. Podríamos revisarlo afirmando que, cuando los cálculos arrojen estos resultados contradictorios, ningún resultado sería peor que los otros. Este principio revisado implicaría que algunos resultados serían peores que otros, pero proporcionaría sólo una ordenación muy parcial, sin ser capaz de apoyar las respuestas que queremos en muchos casos. (Aquí y en otros lugares estoy en deuda con la correspondencia mantenida con M. Woodford.)

Ahora hemos aprendido que la apelación a esta Restricción fracasa, y que es más difícil explicar por qué deberíamos rechazar el Principio Total. Supongamos en primer lugar que, si causamos que exista alguien que tendrá una vida de las que valen la pena vivirse, con ello no beneficiamos a esta persona. Si esto es así, no podemos apelar a la Restricción de las Personas Afectadas, puesto que entonces seríamos incapaces de resolver el Problema de la No- Identidad. Según esta alternativa, no podemos rechazar el Principio Impersonal del Total con la afirmación de que, como no se refiere a las personas afectadas, toma la forma incorrecta —la del Modelo de Producción de Vapor—. Para resolver el Problema de la No-Identidad tenemos que apelar a un principio que no se refiera a las personas afectadas.

Supongamos después que, al causar que alguien exista, podemos con ello beneficiar a esta persona. Si esto es así, podemos recurrir a la Restricción de las Personas Afectadas. Pero esto no consigue nada. El Principio Amplio del Total *vuelve a exponer el Principio Impersonal en una forma de personas afectadas.* Para evitar la Conclusión Repugnante, tenemos que afirmar que deberíamos rechazar este principio. Y es más difícil explicar por qué deberíamos hacerlo. Como este principio se refiere a las personas afectadas, no podemos decir que toma la forma errónea, tratando a la gente como meros contenedores o productores de valor. Es más fácil negar que sería mejor que hubiera más felicidad, o más de todo lo que hace la vida digna de vivirse. Es más difícil negar que sería mejor que la gente fuese más beneficiada.

Aunque es más difícil negar el Principio Amplio del Total, no estamos forzados a aceptarlo. Es defendible negar una de las afirmaciones de este principio: la de que causar que se exista puede beneficiar. Y, aunque creamos que causar que se exista puede beneficiar, hay al menos otro principio que se refiere a las personas afectadas: el Principio Amplio de la Media. Este principio no implica la Conclusión Repugnante. Si lo que sucediera fuese Z en vez de A, esto sería «peor para las personas» en el sentido medio amplio. Las personas-Z recibirían, comparadas con las personas-A, un beneficio más pequeño por persona.

El Principio Amplio de la Media vuelve a exponer en términos de personas afectadas el Principio Impersonal de la Media. Necesitamos un nuevo principio de beneficencia. Necesitamos un principio que resuelva el Problema de la No-Identidad y a la vez evite la Conclusión Repugnante. También es deseable que explique la Asimetría. Tanto en su forma impersonal como en su forma de personas afectadas, el Principio de la Media logra los dos primeros objetivos, y en parte también el tercero [35]. ¿Es este principio lo que queremos: la mejor explicación de la beneficencia?

137 Teorías posibles

El Principio de la Media es sólo uno de los que logra nuestros objetivos. Será más fácil juzgar este principio cuando yo haya descrito algunas de estas alternativas.

No podemos rechazar el Principio Amplio del Total con la afirmación de que adopta la forma incorrecta. Pero podemos afirmar que este principio da una respuesta incorrecta a una pregunta que planteé arriba. Pregunté, «¿Cuáles son los valores relativos, durante un período determinado, de la *calidad* de vida y de la *cantidad* tanto de la felicidad como de todo lo demás que hace la vida digna de vivirse?». De la variedad de posibles respuestas, el Principio del Total está en un extremo. Según este principio, tanto en su forma impersonal como en su forma referida a personas afectadas, *sólo la cantidad* tiene valor.

Según un modo de ver las cosas menos radical, la calidad y la cantidad tienen valor las dos. Según este modo de ver las cosas, si la cantidad de la felicidad es la misma, será peor si esta felicidad entra en más vidas de una calidad más baja. Para que B no sea peor que A, quizás tenga que contener, no al menos la misma suma total de felicidad, sino una suma mayor, quizás dos veces más, o

[35] Ver la discusión de este principio que hace McMahan en Mcmahan (1).

diez veces más, o incluso más. Una afirmación parecida podría hacerse sobre la cantidad de todo lo que hace la vida digna de ser vivida.

El otro extremo es la idea de que *sólo la cualidad* tiene valor. Según este modo de pensar, si la gente resulta menos favorecida, esto *nunca* puede ser moralmente compensado por el hecho de que haya más gente viviendo. Una pérdida en calidad nunca puede ser compensada por una ganancia en cantidad. El Principio de la Media es una versión de esta concepción.

Podría mantenerse esta concepción, no sólo en relación con diferentes poblaciones posibles durante un período de tiempo, sino también en relación con la cuestión más amplia de cuántas personas debería haber siempre. Como dije antes, la Tesis de la Media debe adoptar una forma temporalmente neutral. En esta forma, contesta a esta pregunta más amplia.

Supongamos que el número de los que alguna vez vivieron hubiera sido muy pequeño. En una historia imaginaria, Eva y Adán tienen vidas que son perfectamente dignas de ser vividas, y no tienen hijos. Son las únicas personas que jamás vivirán. En una historia posible diferente, millones de personas tienen vidas que son perfectamente dignas de ser vividas, pero ninguna tiene una vida que sea tan buena como las que Eva y Adán habrían tenido. En esta segunda historia posible, la calidad de vida es ligeramente más baja, pero hay muchísima más cantidad tanto de felicidad como de todo lo demás que hace la vida digna de ser vivida. Según la Tesis de la Media, o la afirmación más simple de que sólo importa la calidad, la primera historia posible sería mejor. Hay algunas personas que aceptan esta conclusión.

Otros aceptarían versiones menos radicales de esta afirmación. Como queremos evitar la Conclusión Repugnante, ayudará distinguir dos grupos de respuestas, que pueden ser presentadas con dos preguntas, como se muestra en la página 691.

(1) es el Principio Total. El Principio de la Media es una versión de (7). Podemos permitirnos ignorar aquí las diferencias entre algunas de estas concepciones. Pienso que (2) es más plausible que (1). Y pienso que (5) es más plausible que (6), que es más plausible que (7). Pero, para nuestros propósitos, podemos ignorar estos detalles [35ᵇ].

Podemos ignorar la diferencia entre (1) y (2), puesto que ambas concepciones responden Sí a mi primera pregunta. Las dos afirman que, como no hay límite superior para el valor de la cantidad, cualquier pérdida en la calidad de vida podría ser moralmente compensada por una ganancia suficiente en cantidad. Las dos concepciones implican, por tanto, la Conclusión Repugnante. La diferencia es sólo que, según la concepción (2), la población de Z tiene que ser más grande. Estas son variantes de la idea de que *la cantidad siempre podría tener más peso que la calidad*.

De (3) a (7) son variantes de la negación de esta idea. De acuerdo con todo lo que va de (3) a (7), la cantidad no puede tener siempre más peso que la calidad.

[35ᵇ] La concepción (5) me fue sugerida por J. McMahan. Y los dos estamos influenciados por Hurka (3). La concepción (1) difiere de todas las demás de la siguiente manera. De acuerdo con (1), el valor de un resultado es la suma de su valor para todas las personas en este resultado. Si alguien lleva una vida digna de vivirse, esta vida tiene valor para esta persona; y tiene más valor si es más digna de vivirse. Llamo a esta clase de valor *personal*. Según la concepción (1), el mejor resultado es aquel que tiene más valor personal.

Según las otras seis concepciones, esto se niega. Uno de dos resultados puede ser peor aunque tenga más valor personal. Lo cual puede parecer inverosímil. Si un resultado tiene más valor para las personas en este resultado, esto puede parecer una poderosa razón para juzgar que es el mejor resultado. Según las concepciones de la (2) a la (7), el valor de un resultado no se corresponde con su valor personal. En este sentido, tales concepciones son *impersonales*. Lo cual puede parecer una objeción contra estas concepciones.

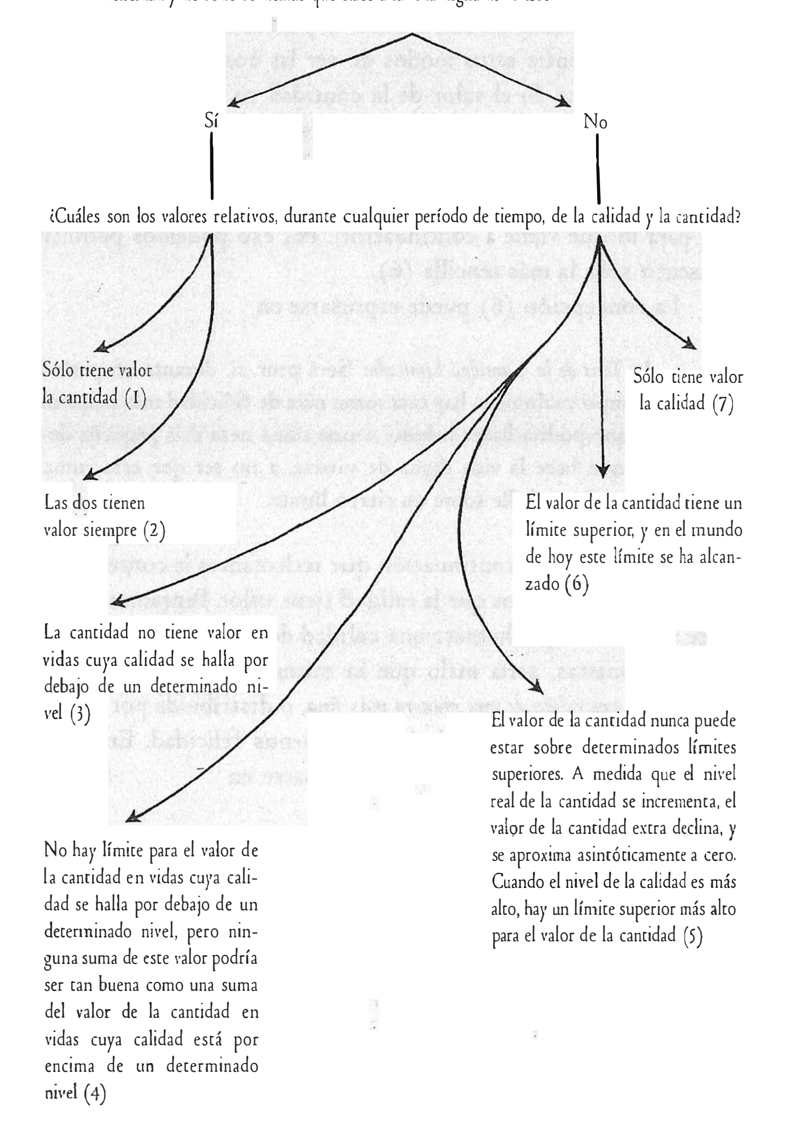
Aunque en este sentido sólo (1) es una concepción personal, en otro sentido (1) es la única concepción impersonal. (1) es la única que no se preocupa por la calidad de las vidas de las personas, o por la cantidad de felicidad que hay en esas vidas. De acuerdo con (1), si la cantidad total es la misma, no representa ninguna diferencia moral el que la calidad de las vidas de las personas sea mucho más baja. Todas las otras concepciones están interesadas en la calidad de las vidas de las personas. *Por esta razón* son estas concepciones, en el primer sentido, impersonales: porque niegan que el valor de un resultado corresponda a la suma total de su valor personal. Como niegan esta afirmación porque están interesadas en la calidad de las vidas de las personas, no supone ninguna objeción contra estas concepciones el hecho de que, en el primer sentido, sean impersonales.

Podemos ignorar la diferencia entre (5) y (6). Podemos considerarlas como variantes de la idea de que *el valor de la cantidad tiene límites superiores*. (7) es la variante en que este límite superior es siempre cero. De (5) a (7) están en desacuerdo sólo cuando es pequeña la población mundial. (5) y (6) afirman entonces lo que niega (7), que habría algún valor en la mayor cantidad.

Según la concepción (5), para cualquier nivel de calidad, el valor en la cantidad extra se aproxima asintóticamente a cero. Esto significa que el valor de cada unidad extra de cantidad se acerca cada vez más a cero, pero sin alcanzarlo realmente. Esta afirmación no implica que el valor de la cantidad tenga límites superiores. Supongamos que el valor de cada vida extra digna de ser vivida fuese como sigue: 1, 1/2, 1/3, 1/4, 1/5, 1/6, etc. El valor de cada vida extra se aproxima aquí a cero; pero no hay límite para el valor de tal secuencia. Como (5) afirma que tales límites existen, el valor de cada vida extra tiene que disminuir de un modo diferente. Un ejemplo sería: 1, 1/2, 1/4, 1/8, 1/16, etc. Aunque siempre habría algún valor en cada vida extra, esta secuencia tiene un límite superior: 2. Pero, con un número finito de vidas extra, este límite superior nunca se alcanza del todo.

La población real del mundo en la actualidad consta de varios miles de millones de personas, la mayoría de las cuales creen que sus vidas son dignas de vivirse. La mayoría de los que aceptan la concepción (5) estarían de acuerdo en que, en el mundo actual, el valor de la cantidad extra puede considerarse que ha alcanzado su límite. Por eso supuse que, en mi resultado imaginario A, habría diez mil millones de personas

Si hay una calidad de vida más baja, ¿podría esto ser siempre compensado moralmente por una ganancia suficiente en el número de personas que existen, y tienen vidas dignas de vivirse? Más precisamente, una pérdida de calidad durante un período de tiempo, ¿podría compensarse con una ganancia suficiente en la cantidad de felicidad y de todo lo demás que hace a la vida digna de vivirse?



viviendo. Esto hace a mi ejemplo relevantemente similar al mundo real; y lo convierte en un ejemplo en que las concepciones de la (5) a la (7) están de acuerdo.

En lo que sigue, también ignoraré la diferencia entre (5) y (6). (5) es más plausible. Pero, cuando la población real es muy grande, la diferencia entre estos modos de ver las cosas tiene poca significación práctica. Si el valor de la cantidad se haya extremadamente próximo a su límite superior, puede considerarse que ha alcanzado este límite. Al afirmar que hay un solo límite superior, (6) es de nuevo menos plausible que (5). Pero esta diferencia no es relevante para lo que viene a continuación. Por eso podemos permitirnos discutir sólo la más sencilla (6).

La concepción (6) puede expresarse en

> *La Tesis de la Cantidad Limitada*: Será peor si, durante un período de tiempo cualquiera, hay una suma neta de felicidad más pequeña que la que podría haber habido, o una suma neta más pequeña de todo lo que hace la vida digna de vivirse, a no ser que esta suma más pequeña se halle sobre un cierto límite.

Supongamos a continuación que rechazamos la concepción (1), puesto que pensamos que la calidad tiene valor. Pensamos que siempre sería malo que hubiese una calidad de vida más baja. En términos hedonistas, sería malo que la misma suma total de felicidad estuviera *extendida de una manera más fina*, o distribuida por más vidas, de forma que en cada vida hubiera menos felicidad. En su forma más simple, esta creencia puede expresarse en

> *La Tesis de la Calidad en Dos Niveles*: Será peor si los que viven van a resultar *todos* menos favorecidos, o con una calidad de vida más baja, que *todos* los que habrían vivido.

Según esta tesis, si en uno de dos resultados todos resultaran menos favorecidos, esto sería un mal rasgo de este resultado. No se sigue que, de los dos, este sea el peor resultado. Este mal rasgo podría compensarse con un buen rasgo. Como la Tesis de la Calidad en Dos Niveles es simplemente una afirmación sobre la maldad de un rasgo, creo que la mayor parte de nosotros la aceptaría.

Supongamos que aceptamos tanto esta como la Tesis de la Cantidad Limitada. En casos que implican poblaciones pequeñas, pensaremos que la calidad más baja podría ser compensada por la mayor cantidad. Pero, si el valor de la cantidad puede considerarse que ha alcanzado su límite, pensaremos que sólo la calidad tiene valor. Como he dicho, esto es lo que la mayoría de nosotros piensa del mundo real en el siglo XX. Y esto es lo que pensaríamos en relación con mis imaginarios resultados A y B. Como creeríamos que, en A, el valor de la cantidad puede considerarse que ha alcanzado su límite superior, pensaríamos que B sería peor que A. Esto sería así por mucho que B contuviera muchas más personas. Cuando consideramos poblaciones tan grandes como diez mil millones, pensaremos que una pérdida en calidad nunca podría ser moralmente compensada por una ganancia en cantidad. Sería siempre peor si, en vez de un grupo de diez mil millones de personas, hubiera más personas que resultaran *todas* menos favorecidas. Esto sería peor aunque, en ambos resultados, la calidad de vida fuese muy alta. [Si rechazamos esta última afirmación, deberíamos volver a la concepción (3), discutida en la Sección 139.]

La Tesis de la Calidad en Dos Niveles no incluiría muchos casos reales. Rara vez sería verdadero que, en uno de dos resultados que fuesen producidos por dos ritmos de crecimiento demográfico, *todos* resultaran menos favorecidos que *todos* en el otro resultado. Si pensamos que la calidad tiene valor, tenemos que hacer alguna otra afirmación para incluir la mayoría de los casos reales. Una afirmación semejante es la Tesis de la Media, bien en su forma de personas afectadas, bien en su forma impersonal. Muchos apelan a esta concepción. Pero, como defenderé, deberíamos rechazar la Tesis de la Media. (Los que apelan a esta concepción pueden haber aceptado la Tesis de la Calidad en Dos Niveles, y luego, temerariamente, haber cambiado «todos» a «por término medio».)

Para incluir la mayor parte de los casos reales, necesitaríamos una afirmación más complicada. Pero podemos ignorar estas complicaciones. Será suficiente con considerar las dos opiniones que acabo de describir. Una es la de que el valor de la cantidad tiene,

dentro de cualquier período de tiempo, un límite superior. La otra es la de que la calidad tiene valor.

Estoy buscando la Teoría X, la nueva teoría sobre la beneficencia que resuelve el Problema de la No-Identidad y a la vez evita la Conclusión Repugnante. Más en general, la Teoría X sería la mejor teoría de la beneficencia. Tendría implicaciones aceptables cuando se aplicase a todas las elecciones que hacemos, incluyendo las que afectan tanto a las identidades como al número de las personas futuras. En una teoría moral completa, no podemos evitar la cuestión de cuántas personas debería haber siempre. Pero he pospuesto esta cuestión. Estoy discutiendo qué deberíamos pensar de los diferentes resultados durante períodos de tiempo determinados.

Tal vez esperemos que la Teoría X sea sencilla. Si debemos aceptar la Conclusión Repugnante, esta teoría podría ser simple: podría ser la afirmación de que sólo la cantidad tiene valor. Todas las versiones del Principio Total dan expresión a esta afirmación. Pero me parece que deberíamos rechazar la Conclusión Repugnante. Partiendo de esta base, podríamos apelar a otra teoría sencilla: la afirmación de que sólo la calidad tiene valor. El Principio de la Media es una versión de esta afirmación. Pero creo que, como su opuesta, esta afirmación es demasiado radical. Considero que la mejor teoría de la beneficencia tiene que afirmar que ambas, la calidad y la cantidad, tienen valor.

Si estas creencias están justificadas, la mejor teoría podría ser la combinación de las dos opiniones que acabo de describir. Según esta teoría, la calidad siempre tiene valor, y la cantidad tiene valor de modo que, en cualquier período de tiempo, no puede estar sobre algún límite superior. Esta teoría logra, en mis ejemplos simplificados, los dos resultados que he estado tratando de lograr. Resuelve el Problema de la No-Identidad y evita la Conclusión Repugnante. Si esta teoría es defendible, podría ser una versión simplificada de X: la mejor teoría de la beneficencia.

¿Es defendible esta teoría? No en su forma presente.

138. La suma de sufrimiento

Volvamos a considerar

> *Los Dos Infiernos*. En el *Infierno Uno*, la última generación consta de diez personas inocentes, cada una de las cuales sufre un gran dolor durante cincuenta años. La vida de estas personas es mucho peor que nada. Todos se darían muerte a sí mismos si pudieran. En el *Infierno Dos*, la última generación consta no de diez sino de diez millones de personas inocentes, cada una de las cuales sufre un dolor igual de intenso durante cincuenta años menos un día.

Cuando consideramos estos Infiernos imaginarios, no podemos apelar sólo a la Tesis de la Calidad en Dos Niveles, o a la Tesis de la Media. Según estas concepciones, el Infierno Uno sería peor, puesto que las vidas de diez personas serían todas ellas un poco peores que las vidas de diez millones de personas. Podemos conceder que, en este aspecto, el Infierno Uno sería peor. Sería malo que, en este Infierno, cada una de las personas que existe tenga que soportar un sufrimiento un poco mayor. Pero puede decirse que, en otro aspecto, el Infierno Dos sería peor. En este Infierno, la suma total de sufrimiento es casi un millón de veces mayor. Y puede decirse que este vasto incremento en la suma de sufrimiento pesa más moralmente que la muy pequeña reducción en la suma de sufrimiento dentro de cada vida.

¿Podemos negar estas afirmaciones? Tendríamos que apelar al

> *Principio del Sufrimiento Limitado*: Será malo si, en cualquier momento dado, hay una suma de sufrimiento mayor que la que podría haber habido, a no ser que esta suma mayor se halle sobre un cierto límite.

Según esta opinión, a medida que la suma de sufrimiento aumenta, el sufrimiento extra importa menos. Y este disvalor se aproxima a cero. Más allá de cierto punto, para propósitos prácticos, el sufrimiento extra deja de importar. Si más personas tienen que soportar incluso el más extremo dolor, esto no sería malo en absoluto.

Esta concepcion es muy inverosímil, mucho más inverosímil que la Conclusión Repugnante. Cuando consideramos la maldad del sufrimiento, deberíamos afirmar que esta maldad no tiene límite superior. Siempre es malo que una persona más tenga que soportar un dolor extremo. Y es siempre igual de malo, por mucho que tengan vidas parecidas muchos otros. La maldad del sufrimiento extra nunca disminuye.

En el caso del sufrimiento, no hay límite superior para el disvalor de la cantidad. ¿Pensamos, cuando consideramos el sufrimiento, que sólo importa la cantidad? La mayoría de nosotros no. Pensaríamos que sería mejor si la misma cantidad de sufrimiento fuese distribuida de una manera más fina, en un número mayor de vidas. Esto lo implica, en los casos sencillos, la Tesis de la Calidad en Dos Niveles.

Deberíamos añadir estas afirmaciones sobre el sufrimiento a la teoría que describí. Nuestra teoría combina ahora tres ideas. Pensamos que la calidad siempre tiene valor: que sería siempre malo que la gente resultase menos favorecida de lo que podría haberlo sido. Pensamos también que hay valor en la cantidad de felicidad, y de todo lo demás que hace la vida digna de vivirse, pero que este valor tiene, en cualquier período de tiempo, un límite superior. Y ahora pensamos que hay disvalor, o maldad, en la cantidad no sólo de sufrimiento sino de todo lo demás que hace la vida *digna de terminar*. Y pensamos que, para esta clase de maldad, no hay límite.

Esta nueva teoría implica

> *La Conclusión Ridícula*: Si hubiera diez mil millones de personas vivas, todas con una calidad de vida más o menos como la calidad media de las vidas que vive la presente población del mundo, tiene que haber una población imaginable mucho mayor cuya existencia sería *peor*, aunque *todos* sus miembros tuvieran una calidad de vida *mucho más alta*. Esta población mucho más grande estaría peor porque, en cada una de estas vidas, habría un sufrimiento intenso [36].

[36] Desarrollo un argumento aportado por R. Sikora en Sikora y Barry.

Incluso en vidas de una calidad mucho más alta que la nuestra, podría haber un intenso sufrimiento. En la imaginaria población mayor, todo el mundo tiene una vida así. Estas vidas serían perfectamente dignas de ser vividas porque este sufrimiento sería ampliamente compensado por la felicidad, y por las demás cosas que hacen la vida digna de vivirse. Pero, según nuestra nueva teoría, este sufrimiento no puede ser *moralmente* compensado por estas otras cosas. Dado el tamaño de la población, y su alta calidad de vida, no hay ningún valor positivo en una cantidad extra. El sufrimiento en cada vida extra sería un mal rasgo, que haría el resultado peor. Según nuestra teoría, este mal rasgo no puede ser moralmente compensado por la felicidad mucho mayor en cada vida extra, o por las otras cosas que hacen a estas vidas muy dignas de ser vividas. Estos no son buenos rasgos, que hagan mejor el resultado, puesto que el valor positivo de la cantidad ya ha alcanzado su límite.

Según nuestra teoría, cualquier vida extra tendría sólo dos rasgos moralmente relevantes. Uno sería la calidad de esta vida, el otro el sufrimiento que contiene. La Conclusión Ridícula compara dos poblaciones posibles. Una contiene diez mil millones de personas, todas las cuales tienen una calidad de vida que es más o menos como la calidad media de las vidas que ahora se están en realidad viviendo. La otra población sería mucho más grande, y todos sus miembros tendrían una calidad de vida mucho más alta. (Podemos suponer que la mayor parte de esta población mucho mayor vive en muchos planetas similares a la Tierra que han venido a parar al Sistema Solar, y ahora forman parte de él.) Comparada con la existencia de la población más pequeña, la existencia de esta población más grande sería de un modo mejor, y de otro peor. Este resultado tendría el buen rasgo de que todo el mundo tiene una calidad de vida mucho más alta. Tendría el mal rasgo de que habría una suma mayor de sufrimiento intenso. Si, con la misma calidad de vida, imaginamos que esta población es aun mayor, esto no haría al buen rasgo mejor. Pero haría al mal rasgo peor, dado que habría una suma aun mayor de intenso sufrimiento. Según nuestro modo de pensar, no hay límite para la maldad de una suma mayor de un sufrimiento semejante. Como la maldad de este rasgo no tiene límite, con una

población lo suficientemente grande tiene que pesar más que el buen rasgo. Nuestra nueva teoría implica, evidentemente, la Conclusión Ridícula.

¿Cómo podemos evitar esta conclusión? No podemos de una manera verosímil colocar un límite superior sobre la maldad de la suma de sufrimiento. Pero podemos distinguir dos clases de sufrimiento. El sufrimiento es *compensado* si está dentro de una vida que vale la pena vivirse. Si está dentro de una vida que no vale la pena vivirse es *no compensado*. Podemos afirmar que, de estas dos clases de sufrimiento, sólo la segunda es mala. Supongamos que, próximo al final de una vida que ha sido perfectamente digna de vivirse, alguien empieza a sufrir intensamente. No podemos afirmar, de manera justificable, que, cuando esta persona está sufriendo intensamente, esto no sea malo. Pero podríamos decir

> Es siempre malo que haya sufrimiento no compensado. Para esta maldad no hay límite superior. Y si sufre más una persona, y tiene una vida que no vale la pena vivirse, esto es siempre igualmente malo. La maldad de este sufrimiento no puede reducirse por el hecho de que otra personas sean felices. Cuando consideramos la maldad del sufrimiento en una vida que no merece la pena vivirse, es irrelevante lo que les ocurre a las demás personas [37].
>
> Hay dos modos en que podría haber más sufrimiento compensado: (1) Podría haber más sufrimiento en una vida que ahora se está viviendo y que vale la pena vivirse. (2) Podría haber una persona extra que existe, con una vida que vale la pena vivirse, pero que contiene algún sufrimiento. De estos dos, sólo (1) es malo. Por eso podemos rechazar la Conclusión Ridícula. No sería un mal rasgo que, en la imaginaria población mayor, hubiera un sufrimiento más intenso. No sería un rasgo malo porque este sufrimiento entraría en vidas extra que valen la pena vivirse. No sería peor que esta gente extra existiera. Lo que decimos de este sufrimiento extra es, no que ser malo, sino que habría sido mejor si estas vidas no lo hubieran contenido.

[37] Es plausible apelar aquí, como hacen los no utilitaristas, a la condición separada de las personas. Lo que ocurre en otras vidas de ningún modo puede compensar, o reducir la maldad del sufrimiento en cada una de estas vidas.

Puede objetarse: «si este sufrimiento extra no es malo, ¿por qué habría sido mejor que estas vidas no lo hubieran contenido?». Podríamos contestar, «Porque esto habría hecho la calidad de vida aun más elevada».

Puede objetarse a continuación: «si niegas que el sufrimiento extra sea malo cuando entra en vidas extra que vale la pena vivir, tienes que estar asumiendo implícitamente que estas vidas contienen rasgos buenos que pesan más que la maldad del sufrimiento. Esto socava tu afirmación de que el valor positivo de la cantidad ha alcanzado su límite superior».

Podríamos contestar por nuestra parte: «Nuestra concepción distingue valor *personal* y valor *moral*. Lo que llamamos el valor de una vida no es su valor personal —su valor para la persona cuya vida es— sino el valor que esta vida aporta al resultado. Cuando decimos que una vida no tiene valor, esto significa que vivir esta vida no hace el resultado mejor. En otras palabras, no habría sido peor en sí mismo que la persona con esta vida no hubiera existido nunca. Si esto hubiera sido peor lo habría sido sólo a causa de sus efectos en otras personas. Dada esta distinción, podemos responder a tu objeción. El sufrimiento en estas vidas extra tiene disvalor personal, o sea, es malo para las personas que vivan estas vidas. Este disvalor queda compensado, para estas personas, por el valor personal de su felicidad, y de las otras cosas que hacen que sus vidas sean dignas de vivirse. Este valor personal no es valor moral. La existencia de estas personas extra no hace el resultado mejor. De una manera parecida, dado que su sufrimiento es compensado, o contrarrestado, por el valor personal de las mismas vidas, el disvalor personal de este sufrimiento extra no tiene disvalor moral. Cuando hay más sufrimiento sólo porque hay más vidas vividas que vale la pena vivir, este sufrimiento extra no hace el resultado peor» [38].

Consideremos a continuación tres poblaciones posibles, durante un período determinado. La primera es la población actual del

[38] Para una distinción similar entre valor personal y valor moral, véase Nagel (3), pp. 97-139.

mundo, durante el último cuarto del siglo XX. La segunda es una población mayor, en que *casi* todos sus integrantes tienen una calidad de vida mucho más alta. Como antes, la mayor parte de los que la integran viven en otros planetas semejantes a la Tierra, que han llegado a formar parte del Sistema Solar.

En esta población mayor, hay algunas personas desafortunadas que sufren y cuyas vidas no vale la pena vivirlas. Podrían ser como el Niño Desgraciado descrito arriba, ese que está tan enfermo que nunca llega a desarrollarse, vive sólo unos pocos años, y sufre un dolor que no puede ser aliviado del todo. O bien podrían estar aquejados de cierta enfermedad mental que hace sufrir mucho y que dura toda la vida. Estas personas desafortunadas son, en proporción, muy pocas. En esta población imaginaria habría una de tales personas por cada diez mil millones.

Muchos piensan que sería incorrecto llevar la felicidad a millones torturando a un niño inocente [39]. Aunque las proporciones sean similares, mi caso imaginario es muy diferente. Es por pura mala suerte que, por cada diez mil millones en esta población mayor, haya una persona con una enfermedad que le hace sufrir, y hace que su vida no sea digna de ser vivida. Estas personas desafortunadas no se suicidan o bien porque no se desarrollan lo suficiente, o bien a causa de la naturaleza de su enfermedad mental. Algunos de nosotros pensaríamos que, por su propio bien, se debería matar a tales personas. Pero este no es el modo de pensar de mi gente imaginaria. Aunque piensan que sería incorrecto matar a los pocos desafortunados, hacen todo lo que pueden para aliviar su sufrimiento.

Supongamos a continuación que, porque sólo hay una de tales personas por cada diez mil millones, habría *menos* de estas personas en esta población mayor que las que hay en el mundo real ahora. Esto se puede suponer de forma verosímil aunque esta población

[39] Consideremos el discurso del Gran Inquisidor en *Los hermanos Karamazov* de Dostoievsky, y la afirmación de R. B. Perry, citada arriba, de que «la felicidad de un millón por alguna razón no puede en absoluto... mitigar siquiera la tortura de uno».

imaginaria tenga muchas veces el tamaño de la población mundial presente. En la población actual del mundo hay algunas personas cuyas enfermedades les hacen sufrir, y hacen que sus vidas no valga la pena vivirlas. Sería difícil calcular la proporción que hay entre tales personas y las personas reales cuyas vidas sí que valen la pena. Pero queda claro que esta proporción es *mucho* más alta que uno entre diez mil millones.

Comparada con la población real del mundo, esta población imaginaria mayor es de dos modos mejor, y de ninguno peor. La calidad de vida de la mayoría de la gente es mucho más alta. Y hay *menos* sufrimiento no compensado. Comparada con la existencia de la población actual del mundo, la existencia de esta población imaginaria sería evidentemente mucho mejor.

Consideremos ahora una segunda población imaginaria. Es exactamente como la primera, pero muchísimo mayor. Esta población viviría en muchísimos planetas semejantes a la Tierra. Comparada con la existencia de la población real del mundo, la existencia de esta segunda alternativa imaginaria sería en un sentido mejor, y en otro peor. El sentido en que sería mejor es que casi todo el mundo tiene una calidad de vida muchísimo más alta. El sentido en que sería peor es que habría *más* sufrimiento no compensado.

Como antes, si imaginamos que esta población es aún mayor, no hay manera en que esto sería mejor, y una en que sí sería peor. La calidad de vida sería la misma, y el valor positivo de la cantidad habría alcanzado su límite. Si esta población es aún mayor, sus buenos rasgos no serían mejores. Pero su mal rasgo sería peor. Habría aun más sufrimiento no compensado. Y para la maldad de este rasgo no hay límite.

Según nuestra teoría, los buenos rasgos de esta población tienen un valor limitado. Pero no hay límite para el disvalor del rasgo malo. Cada vez que imaginamos que la población es mayor, los rasgos buenos no se harían mejores, pero el mal rasgo se haría siempre peor. Si esta población fuera suficientemente grande, su existencia sería peor que la de la población real del mundo. Y, si esta población fuera suficientemente grande, su rasgo malo pesaría más que sus rasgos buenos. La maldad que podría ser ilimitada tiene que

poder pesar más que la bondad limitada. Si esta población fuera suficientemente grande, su existencia sería intrínsecamente mala. Sería mejor si, durante este período de tiempo, nadie existiera.

Podríamos decir: «Aunque sería mejor en sí mismo que nadie existiera durante este período, esto sería peor vistas las cosas en conjunto, puesto que entonces no habría personas futuras».

Esto no es una respuesta. Consideremos dos historias futuras posibles. En la primera, no hay personas futuras. En la segunda, hay siempre una población de la clase que acabo de describir. Como la existencia de semejante población sería, según nuestra concepción, intrínsecamente mala, tenemos que concluir que sería mejor que no hubiera personas futuras.

Dada nuestra teoría, estamos obligados a aceptar

> *La Conclusión Absurda*: En un resultado posible, existirían durante un siglo futuro tanto una población en la Tierra que es como la población real presente de la Tierra, como un número enorme de otras personas, viviendo en planetas semejantes a la Tierra que han llegado a formar parte del Sistema Solar. Casi todas las personas de estos otros planetas tendrían una calidad de vida muy por encima de la que disfruta la mayoría de la población real de la Tierra. Por cada diez mil millones de estas otras personas habría un desafortunado con una enfermedad que le hace sufrir y llevar una vida que no es digna de vivirse.
>
> En un segundo resultado posible, habría el mismo número enorme de personas futuras extra, con una calidad de vida igual de alta para todos salvo un desafortunado cada diez mil millones. Pero este número enorme de personas futuras extra no vivirían todas en un siglo futuro. Cada diez mil millones de estas personas viviría en cada uno de muchísimos siglos futuros.
>
> Según nuestra concepción, el primer resultado sería muy malo, mucho peor que si no hubiera nadie de estas personas futuras extra. El segundo resultado sería muy bueno. El primero sería muy malo y el segundo muy bueno aunque en ambos hubiera el mismo número de personas futuras extra, con la mismísima alta calidad de vida para todos salvo un desafortunado por cada diez mil millones.

Puede que esta conclusión no sea ridícula. Pero su carácter absurdo no se nos oculta. El primer resultado es exactamente igual

que el segundo, salvo que todas las personas futuras extra viven en el mismo en vez de en diferentes siglos. Si el segundo resultado fuese muy bueno, y el primero se diferencia sólo en lo que se refiere a *cuándo* vive la gente, ¿cómo puede esta diferencia en posición temporal hacer al primer resultado muy malo?

Hay una manera en que una diferencia en posición temporal podría traducirse en una diferencia moral. La desigualdad dentro de una generación puede ser peor que la desigualdad entre diferentes generaciones. Pero, en los dos resultados que estamos considerando, la desigualdad sería la misma. Sería verdadero en ambos que, por cada diez mil millones de personas extra, una persona resulta mucho menos favorecida que todas los demás.

Podemos suponer que, en los dos resultados, *todas las vidas que se viven siempre serían idénticas, tanto en número como en calidad*. En estos aspectos los resultados son exactamente iguales. Pero, a causa de la diferencia en cuándo se viven estas vidas, concluimos que el primer resultado sería muy malo y el segundo muy bueno.

La Conclusión Absurda está evidentemente implicada en nuestra última concepción. Según ella, el valor positivo de la cantidad tiene un límite superior en cualquier período de tiempo, mientras que no hay tal límite para el valor negativo de la cantidad, o para la maldad de la suma de sufrimiento no compensado. En el segundo resultado, en que cada diez mil millones de personas extra viven en siglos diferentes, el valor positivo de la cantidad excede ampliamente, en cada siglo, su valor negativo. Por eso este resultado es muy bueno. En el primer resultado, en que estas personas extra viven todas en el mismo siglo, es verdadero lo contrario. Como el valor positivo de la cantidad tiene, en este siglo, un límite superior, este valor queda rebasado ampliamente por la maldad de la suma del sufrimiento no compensado. Por eso este resultado es muy malo. El segundo resultado es muy bueno, y el primero muy malo, aunque en ambos las vidas que siempre se viven son idénticas en número y en calidad.

No puedo creerme esta conclusión. Es una conclusión que se desprende de la asimetría en nuestras afirmaciones sobre el valor de

la cantidad. Según nuestro modo de ver las cosas, mientras que no hay límite, en ningún período de tiempo, para el disvalor de la cantidad, sí lo hay para su valor. A fin de evitar la Conclusión Absurda, tenemos que abandonar este modo de ver las cosas.

139. El nivel sin valor

La mayoría de nosotros quiere evitar la Conclusión Repugnante. Un modo de hacerlo es afirmar que hay un límite para el valor positivo de la cantidad. Pero, como acabamos de ver, tenemos que abandonar esta opinión. En la página 691 mencioné dos alternativas: las concepciones (3) y (4).

La Tesis (3) es

> *La Apelación al Nivel Sin Valor*: La cantidad no tiene valor en vidas cuya calidad se halla debajo de un cierto nivel. Si estas vidas son dignas de vivirse, tienen valor personal —valor para las personas cuyas vidas son—. Pero el hecho de que tales vidas sean vividas no hace mejor el resultado.

Podemos hacer esta idea más plausible si añadimos las afirmaciones siguientes. El Nivel Sin Valor no implica una discontinuidad brusca. Si la calidad de vida es muy alta, el valor moral de cada vida —la cantidad por la cual hace mejor el resultado— es su valor personal pleno. A un nivel más bajo, el valor de una vida comienza a ser menor que su valor personal pleno, y a niveles más bajos estos dos valores divergen cada vez más. En el Nivel Sin Valor, aunque las vidas todavía sean dignas de vivirse, el primer valor ha llegado a cero. No hay discontinuidad brusca, desde el momento en que a un nivel ligeramente más alto este valor se aproxima a cero.

Podríamos reducir más todavía la discontinuidad. Podríamos afirmar que hay siempre valor en una vida que es digna de ser vivida, pero que, por debajo de cierto nivel, el valor en la cantidad extra disminuye. Debajo de este nivel, el valor total de la cantidad tiene un límite superior, al que nos aproximaríamos asintóticamente, pero que nunca alcanzaríamos del todo. Estas afirmaciones harían más plausible nuestra concepción. Pero en la práctica supondrían sólo una diferencia insignificante. Si la población es lo suficientemente grande, el valor en la cantidad extra por debajo de este nivel estaría muy próximo a cero. Como la diferencia práctica sería insignificante, discutiré la idea más simple que no hace estas afirmaciones adicionales.

Si recurrimos al Nivel Sin Valor, ¿cuál deberíamos suponer que es este nivel? Este nivel es el borde inferior de una banda ancha en la que el valor personal de una vida diverge cada vez más del valor que esta vida aporta al resultado. Si recurrimos al Nivel Sin Valor, necesitamos decidir dónde se sitúa esta banda ancha sobre la escala de la calidad de vida. Puede que encontremos esta pregunta difícil de responder. Pero quizás sea menos difícil que aceptar la Conclusión Repugnante. Como nos hemos visto obligados a abandonar la afirmación de que el valor de la cantidad tiene un límite superior, la apelación al Nivel Sin Valor puede ser el modo menos inverosímil de evitar la Conclusión Repugnante.

Además de apelar al Nivel Sin Valor, deberíamos seguir manteniendo nuestras otras afirmaciones. Deberíamos seguir afirmando que la calidad tiene valor: que siempre sería peor si hubiera una calidad de vida más baja. Y deberíamos seguir afirmando que, si entran en vidas que no vale la pena vivir, hay disvalor, o maldad, en la cantidad tanto de sufrimiento como de todo lo demás que hace la vida merecedora de final. No hay límite para la maldad del sufrimiento no compensado. Deberíamos añadir ahora la afirmación de que hay valor en la cantidad de felicidad, y de todo lo demás que haga la vida digna de ser vivida, siempre que este valor entre en vidas que se hallan sobre el Nivel Sin Valor. Y deberíamos afirmar que para este valor no hay límite.

Si hacemos nuestra esta nueva teoría, evitaremos la Conclusión Absurda. En el primer resultado que describí, casi todo el mundo tiene una vida que se halla sobre el Nivel Sin Valor. Sólo una de cada diez mil millones de personas tiene una vida que no vale la pena vivirse. El valor positivo de la cantidad pesa aquí más que su valor negativo.

Podría objetarse: «Consideremos dos resultados cuya relación proporcionada es la misma que la relación entre A y B, pero en la

que la calidad de vida es mucho más baja. Comparado con el resultado *J*, existiría en el resultado *K* el doble de personas, que resultarían todas menos favorecidas. Asumamos a continuación que, en ambos resultados, la calidad de vida estuviera por debajo de tu Nivel Sin Valor. Según tu nueva teoría, las vidas vividas en estos resultados no tendrían valor. Aunque tuvieran valor personal, no tendrían la clase de valor que hace mejor un resultado. Si en ninguno de los dos resultados las vidas vividas tuvieran valor, esto implicaría que el resultado J no sería mejor que el resultado K. Pero tú dirías que, puesto que la calidad de vida sería en J más alta, J *sería* mejor que K. Tu nueva teoría se contradice a sí misma».

Podríamos responder, por nuestra parte: «Nosotros afirmamos que, si la calidad de vida se halla por debajo del Nivel Sin Valor, las vidas que son vividas, aunque dignas de vivirse, no tienen valor. Tales vidas no tienen la *clase* de valor que la *cantidad* puede proporcionar. Lo que queremos decir con esto es que, si hubiera personas extra viviendo tales vidas, esto no haría mejor el resultado. Hay valor en la cantidad extra sólo cuando la calidad de vida está sobre el Nivel Sin Valor. Pero el valor en la cantidad extra no es la única clase de valor, o el único rasgo que puede hacer mejor un resultado. Nosotros negamos que sólo la cantidad tenga valor. Para nuestra teoría, también lo tiene la calidad. Pensamos que siempre es bueno que las personas resulten más favorecidas, o tengan una calidad de vida más alta. Como hay dos modos en que un resultado puede ser mejor, nuestra teoría no se contradice a sí misma. El resultado J sería mejor que el resultado K, no porque tuviera más del valor de la cantidad, sino porque la calidad de vida sería más alta».

¿Nos proporciona esta teoría un buen escape de la Conclusión Repugnante? Será más fácil responder a esto cuando haya descrito una alternativa.

140. La concepción léxica

Antes que al Nivel Sin Valor, podríamos apelar a la concepción (4). Esta se parece a la idea de Newman sobre el dolor y el pecado.

Newman afirmaba que, aunque los dos eran malos, ninguna cantidad de dolor podría ser tan mala como la más pequeña cantidad de pecado. Según una concepción semejante, no hay una sola escala de valor. Aunque no haya límite para la maldad del dolor, la maldad ilimitada de esta clase no puede ser tan mala como la maldad limitada de otra clase.

Esta idea es coherente. Es más plausible cuando se aplica, como en el caso de Newman, a cosas que están en dos categorías muy diferentes. ¿Podemos aplicarla a una sola categoría: la de las vidas que vale la pena vivir?

Se puede mantener verosímilmente este modo de ver las cosas cuando comparamos ciertas vidas humanas con las vidas de los animales inferiores. Si fuesen conscientes, las ostras podrían tener vidas dignas de ser vividas. Cuando no los crían intensivamente en granjas, las vidas de los cerdos probablemente sean dignas de ser vividas. Pero podemos afirmar convincentemente que, aunque haya algún valor en el hecho de que estas vidas sean vividas, ninguna cantidad de este valor podría ser tan buena como el valor de la vida de Sócrates.

Si vamos a evitar la Conclusión Repugnante de esta manera, tenemos que hacer una afirmación semejante sobre las vidas que viven todos los seres humanos. En la escala de la calidad de vida tenemos que definir dos nuevos niveles. Llamaremos a las vidas *Mediocres* si están por debajo del nivel inferior, y *Maravillosas* si están por encima del nivel superior. Podríamos apelar ahora a

> *La Concepción Léxica*: No hay límite para el valor positivo de la cantidad. Siempre es mejor que se viva una vida extra digna de ser vivida. Pero ninguna cantidad de vidas Mediocres podría tener tanto valor como una vida Maravillosa.

141. Conclusiones

Tenemos que evitar la Conclusión Absurda. Esta conclusión se seguía de la asimetría en nuestras afirmaciones sobre el valor de la cantidad. Pusimos un límite al valor positivo de la cantidad, dentro

de un cierto período de tiempo, pero no pusimos ninguno a su valor negativo. Para evitar la Conclusión Absurda tenemos que abandonar esta asimetría.

No podemos, de una manera convincente, ponerle un límite al valor negativo de la cantidad. Siempre es malo que haya más sufrimiento no compensado, y esta maldad nunca disminuye. Por eso tenemos que quitarle el límite al valor positivo de la cantidad.

Cuando quitamos este límite, necesitamos un nuevo modo de evitar la Conclusión Repugnante. En mi resultado A habría diez mil millones de personas viviendo, todas ellas con una calidad de vida tan elevada como la de las personas más afortunadas que vivan hoy. En el resultado Z habría una población enorme cuyos miembros tienen vidas que apenas vale la pena vivir. Si hay algún valor en el hecho de que se vivan tales vidas, y el valor de la cantidad no tiene límite, Z podría ser mejor que A. El valor de la cantidad, en A, puede ser muy grande. Pero si el valor de la cantidad en Z no tiene límite alguno, este valor podría ser mayor. Y podría ser lo suficientemente grande como para pesar más que el hecho de que la calidad de vida es mucho más baja. Puede haber, una vez más, un gran disvalor en este hecho. Pero este disvalor podría ser compensado por un valor que no tiene límite alguno. Si el valor de la cantidad en Z no tiene límite alguno, este valor podría pesar más que el hecho de que la calidad de vida sea mucho más baja.

Estas observaciones suponen que hay una única escala de valor. Pero podríamos negar este supuesto. Podríamos apelar a la Concepción Léxica. Según ella, es siempre mejor si se vive una vida extra digna de ser vivida. Admitimos que para el valor de la cantidad en Z no hay límite. Pero afirmamos que ninguna cantidad de *este* valor podría ser tan bueno como el valor de una única vida vivida sobre el Nivel Maravilloso. Podemos suponer que, en A, cada vida se halla por encima de este nivel. Según la Concepción Léxica, sea el que sea el tamaño de la población de Z, Z sería peor que A.

Nuestra alternativa es apelar al Nivel Sin Valor. Entonces afirmaríamos que, para propósitos prácticos, las vidas que son vividas por debajo de este nivel podemos asumir que no tienen valor alguno. Si estas vidas son dignas de ser vividas, tienen valor

personal: pero no hacen mejor el resultado. La cantidad tiene valor sólo en vidas que están por encima del Nivel Sin Valor. Para evitar la Conclusión Absurda, admitimos que para este valor no hay límite.

¿Cuál de estas es la mejor opinión? ¿Deberíamos aceptar la Concepción Léxica, o apelar en cambio al Nivel Sin Valor?

Si apelamos al Nivel Sin Valor, no podremos evitar una variante de la Conclusión Absurda. Tendremos que aceptar

(A) Supongamos que, en una historia del futuro, siempre hubiera un enorme número de personas, y que por cada persona individual que sufre y tiene una vida que no es digna de ser vivida, hubiera diez mil millones de personas cuyas vidas *fuesen* dignas de ser vividas, aunque su calidad de vida no fuera ni con mucho tan alta como el Nivel Sin Valor. Esto sería *peor* que si no hubiera personas futuras.

Tenemos que aceptar (A) puesto que, según nuestra opinión, en este futuro imaginario la cantidad no tendría valor positivo alguno. Habría disvalor en el sufrimiento de estos pocos desafortunados. Si en cambio no hubiera personas futuras, podría tal vez decirse que estábamos ante un caso especial en que habría una pérdida en calidad. Pero la maldad de esta pérdida, aunque grande, podría ser menor que la maldad de la suma de sufrimiento. Esto tiene que ser verdadero con una población lo bastante grande, puesto que esta maldad no tiene límite.

Nuestra nueva teoría implica también

(R) Si hubiera diez mil millones de personas viviendo, todas ellas con una calidad de vida muy alta, tendría que haber una población imaginable mucho mayor cuya existencia sería *mejor*, aunque sus miembros tuvieran vidas que apenas están por encima del Nivel Sin Valor.

Esto es una forma debilitada de la Conclusión Repugnante. Nuestra teoría implica esta conclusión porque implica que, por

encima del Nivel Sin Valor, cualquier pérdida de calidad podría ser siempre moralmente compensada por una ganancia suficiente en cantidad. Esto tiene que venir implicado porque, para evitar la Conclusión Absurda, nuestra teoría deja de afirmar que hay un límite al valor positivo de la cantidad.

(A) es menos absurdo que la Conclusión Absurda. (R) es menos repugnante que la Conclusión Repugnante. Lo poco convincente de estas dos conclusiones depende de dónde situemos el Nivel Sin Valor. Podríamos situar este nivel cerca del nivel donde la vida deja de ser digna de vivirse. Si el Nivel Sin Valor es tan bajo como esto, (A) puede no ser absurda, pero (R) es repugnante. Si subimos el Nivel Sin Valor, (R) es menos repugnante, pero (A) es más absurda.

Si en cambio apelamos a la Concepción Léxica, tendremos que aceptar variantes de (A) y de (R). Estas simplemente sustituyen la palabra «Sin Valor» por la palabra «Mediocre». Explico cómo la Concepción Léxica implica estas conclusiones en la nota [40].

En este capítulo he buscado la Teoría X, la teoría de la beneficencia que resuelve el Problema de la No-Identidad, elude la Conclusión Repugnante, y tiene implicaciones aceptables en todos los casos. Hay un principio que consigue los dos primeros objetivos: el Principio Amplio de la Media de las Personas Afectadas. Este principio afirma que, si se causa que exista alguien con una vida digna de ser vivida, la persona es con ello beneficiada. Como argumento en el Apéndice G, tanto esta afirmación como su negación son defendibles.

[40] Según la Concepción Léxica, tenemos que aceptar una variante de (A), reemplazando «Sin valor» por «Mediocre». La existencia de diez mil millones de personas por debajo de este nivel tendría menos valor que la de una sola persona por encima del Nivel Maravilloso. Si la existencia de estas personas tuviera menos valor que la de esta sola persona, su valor sería más que superado por la existencia de una persona que sufre y tiene una vida que no es digna de vivirse. Según la Concepción Léxica, cuando consideramos vidas *por encima* del Nivel Mediocre, la cantidad siempre podría tener más peso que la calidad.

El Principio Amplio de la Media no puede, por sí mismo, ser la Teoría X. En algunos casos tiene implicaciones inaceptables. Así, implica que el Infierno Uno sería peor que el Infierno Dos aunque, en el Infierno Dos, cada vida es casi tan mala, y la cantidad de sufrimiento no compensado casi un millón de veces mayor. Más en general, el Principio Amplio de la Media no da peso alguno a la suma de semejante sufrimiento. Aunque aceptemos este principio, tenemos que añadir la afirmación de que es siempre malo que haya más sufrimiento no compensado, y que para esta maldad no hay límite.

Cuando se aplica a vidas que son dignas de ser vividas, el Principio Amplio de la Media no da ningún peso a la cantidad de felicidad, y de todo lo demás que hace la vida digna de ser vivida. El principio podría implicar que, en la mejor historia posible, sólo una persona vive. La mayoría de nosotros encontraría esta concepción demasiado radical. La mayoría de nosotros pensaría que hay algún valor en la cantidad, aunque este valor tenga, en un cierto período de tiempo, un límite superior.

El Principio de la Media coincide parcialmente con la más limitada Tesis de la Calidad en Dos Niveles. De acuerdo con esta tesis, cuando comparamos dos resultados A y B, sería peor si todo el mundo en B resultara menos favorecido que todo el mundo en A. Esta tesis incluye pocos casos reales. Pero pronto defenderé que, incluso como parte de una teoría pluralista que apele a varios principios, el Principio de la Media es inaceptable. La Tesis de la Calidad en Dos Niveles tiene que ampliarse de algún otro modo. Si estas afirmaciones están justificadas, todavía no hemos encontrado la Teoría X. Puesto que deberíamos rechazar el Principio de la Media, todavía no hemos resuelto completamente el Problema de la No-Identidad.

También hemos encontrado otro problema. Si nuestra teoría hace las declaraciones que acabo de describir, evita la Conclusión Repugnante a costa de implicar la Conclusión Absurda. Como este es un problema nuevo, hemos hecho otra vez un progreso de tipo negativo. Apelando al Nivel Sin Valor, o a la Concepción Léxica, resolvemos parcialmente este nuevo problema. Simplemente esta-

mos obligados a aceptar tanto (A) como (R), o dos variantes de estas conclusiones. Estas conclusiones son menos repugnantes y menos absurdas. Pero las dos son poco convincentes. No hemos resuelto del todo este problema.

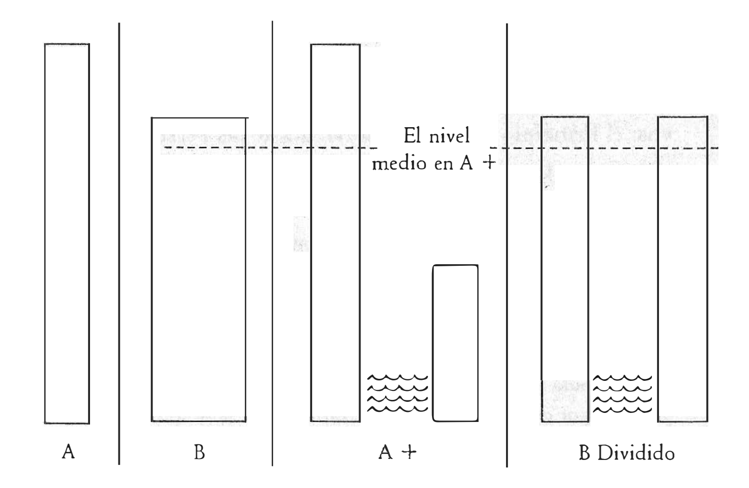
Como argumentaré ahora, hay otro problema.

19

LA PARADOJA DE LA MERA ADICIÓN

142. Mera adición

Consideremos estas alternativas:



Supongamos que recurrimos al Nivel Sin Valor. Asumamos que la calidad de vida en B está por debajo de este nivel. Entonces pen-

saremos que, aunque las personas en B juntas tengan más felicidad que las personas en A, y más de todo lo demás que hace la vida digna de ser vivida, este no es un modo en que B es mejor que A. Las vidas en B están por debajo del nivel en que la cantidad tiene valor. También pensamos que la calidad tiene valor. Creemos que es malo que las personas-B resulten todas menos favorecidas que las personas-A. Como B es de un modo peor que A, y de ningún modo mejor, concluimos que B es peor que A.

Supongamos que recurrimos en cambio a la Concepción Léxica. La calidad de vida en B es alrededor de cuatro quintos de lo que es de alta la calidad de vida en A. Dado este hecho, no podemos asumir verosímilmente ni que las vidas en B se hallen por debajo del Nivel Mediocre ni que las vidas en A se hallen por encima del Nivel Maravilloso. Según nuestra opinión, si B fuese suficientemente grande, sería mejor que A. Una ganancia suficiente en cantidad podría pesar más que su más baja calidad. Pero, como B es sólo el doble de grande que A, podemos asegurar que esto no es así. Podemos concluir una vez más que B es peor que A.

Ahora comparemos A con A +. En A + hay un grupo tan grande como el único grupo que hay en A, y con la calidad de vida igual de alta. A + contiene otro grupo, al que llamo la *gente extra*. Estas personas tienen vidas que valen la pena vivirse, y no afectan a nadie más. La gente extra resulta menos favorecida que la gente del primer grupo. Si esta desigualdad fuera conocida, y fuera además eliminable, podría llevar consigo injusticia social. Por eso asumo, en aras de la simplicidad, que los dos grupos de A + no saben nada de la existencia del otro, y no podrían comunicarse. A + es un posible estado del mundo antes de que se cruzase el Océano Atlántico. A es un estado diferente en el que las Américas están deshabitadas.

¿A + es mejor o peor que A? Será útil definir una nueva expresión. Hay

> *Mera Adición* cuando, en uno de dos resultados, existe gente extra (1) con vidas que valen la pena vivirse, (2) que no afecta a nadie más, y (3) cuya existencia no conlleva injusticia social.

Cuando comparamos A con A +, la Mera Adición en A + baja la calidad media de vida. ¿Es esto un mal efecto, que hace a A + peor que A?

143. Por qué deberíamos rechazar el Principio de la Media

De acuerdo con el Principio de la Media, tanto en su forma de personas afectadas como en su forma impersonal, es peor que haya una calidad media de vida más baja, por vida vivida. Según este principio, A + es peor que A. Pero deberíamos rechazarlo.

Según el Principio de la Media, la historia mejor podría ser una en la que sólo viven Eva y Adán. Sería peor que, en vez de Eva y Adán, vivieran mil millones de millones de otras personas, todas con una calidad de vida que fuese casi tan alta. Aunque esta afirmación es difícil de creer, no es absurda. La segunda historia es de un modo peor. Es malo que la vida de nadie sea tan buenísima como habría sido la de Eva y Adán.

El Principio de la Media tiene otras implicaciones que *son* absurdas. Supongamos que Eva y Adán vivieran esas vidas maravillosas. Según el Principio de la Media sería peor que vivieran mil millones de millones de otras personas, *no en su lugar sino además de ellos*. Sería peor porque bajaría la calidad media de vida. *Este* modo de bajar la media, por Mera Adición, no se puede decir verosímilmente que sea malo.

Una afirmación parecida se aplica al nacimiento de cualquier niño. Que esto sea malo, a los ojos del Principio de la Media, depende de hechos sobre todas las vidas previas. Si los antiguos egipcios tuvieron una calidad de vida muy alta, es más probable que tener un hijo ahora sea malo. Es más probable que el nacimiento de este niño vaya a bajar la calidad media de las vidas que alguna vez se vivieron. Pero la investigación en Egiptología no puede ser relevante para nuestra decisión de si tener o no tener hijos [41].

[41] Aquí me hallo en deuda con Mcmahan (1).

Estas son objeciones a la versión temporalmente neutral del Principio de la Media. ¿Podemos defender alguna otra versión? Ciertos autores asumen que lo que importa es la calidad media de vida de los que vivan después de que actuemos. Pero esto implica el absurdo de que produciríamos un resultado mejor si matáramos a todos excepto a las personas más favorecidas que ahora vivan.

Podría afirmarse que lo que importa es la calidad media de vida de todas las personas que viven en el presente y que vivirán en el futuro. Esta afirmación evita la última implicación. Pero consideremos

> *Cómo Sólo Francia Sobrevive.* En un futuro posible, las personas menos favorecidas del mundo pronto empiezan a tener vidas perfectamente dignas de ser vividas. La calidad de vida en las diferentes naciones sigue entonces subiendo. Aunque cada nación tiene la parte que le corresponde en justicia de los recursos del mundo, cosas tales como el clima y las tradiciones culturales dan a algunas naciones una calidad de vida más alta. Los más favorecidos, durante muchos siglos, son los franceses.

En otro futuro posible, una nueva enfermedad infecciosa hace a

casi todos estériles. Los científicos franceses producen la cantidad justa de un antídoto para toda la población de Francia. Todas las demás naciones dejan de existir. Esto tiene algunos efectos negativos sobre la calidad de vida de los franceses supervivientes. Por ejemplo, ya no hay arte, literatura ni tecnología extranjera nueva que los franceses puedan importar. Estos y otros malos efectos pesan más que cualesquiera efectos buenos. De principio a fin de este segundo futuro posible, los franceses tienen, por consiguiente, una calidad de vida que es ligeramente más baja de lo que sería en el primer futuro posible.

En este segundo futuro la calidad media de vida sería más alta. Los franceses supervivientes tendrían una calidad de vida más baja; y en la época en que la mayoría de la gente se quedó estéril, la mayoría de las vidas de estas personas sería peor. Pero estos dos efectos serían compensados con mucho por la no existencia de las otras naciones del mundo. En el primer futuro, habría miles de millones de personas en estas otras naciones. Estos miles de millones de personas, durante muchos siglos, resultarían menos favorecidas que los franceses en cualquiera de estos futuros. Si estos miles de millones de personas nunca viviesen, la calidad media de las vidas futuras sería así más alta.

Según el Principio de la Media, sería mejor que sólo sobrevivieran los franceses. Esta es otra conclusión absurda. Si estos miles de millones de personas vivieran, sus vidas serían *perfectamente* dignas de ser vividas, y su existencia sería *mejor* para los franceses. Según el Principio de la Media, sería peor que viviesen estas personas, simplemente porque los franceses tienen vidas aun mejores. Como esto es cierto, la existencia de estas personas bajaría la calidad media de vida. Como antes, no se puede pensar verosímilmente que *este* modo de bajar esta media importe.

Una afirmación parecida se aplica otra vez al nacimiento de cualquier niño. Supongamos que, en el futuro lejano, la calidad de vida va a ser durante muchos siglos extremadamente alta. Es entonces más probable que sea malo el que yo tenga un hijo, aun cuando su vida fuese perfectamente digna de vivirse, y su existencia no fuera mala para nadie. Es más probable que la existencia de mi hijo bajara la calidad media de todas las vidas futuras. Esto no puede ser

relevante. Que yo deba tener un hijo no puede depender de cuál vaya a ser la calidad de vida en el futuro distante.

Podríamos revisar el Principio de la Media para evitar esta implicación. Tal revisión conllevaría probablemente trazar una distinción arbitraria [42]. Y la objeción principal todavía se aplica. El Principio de la Media tiene que incluir vidas que coinciden parcialmente. Según este principio, si yo debo tener un hijo todavía depende de hechos irrelevantes acerca de las vidas de otras personas.

Esto queda de manifiesto de la manera más clara en el

> *Infierno Tres.* La mayoría de nosotros lleva vidas que son mucho peor que nada. Las excepciones las constituyen los sádicos tiranos que nos hacen sufrir. El resto de nosotros se suicidaría si pudiera; pero se han arreglado las cosas para que esto sea imposible. Los tiranos

[42] Véase de nuevo Mcmahan (1).

afirman sinceramente que, si nosotros tenemos hijos, ellos les harían sufrir un poco menos.

Según el Principio de la Media, debemos tener estos hijos. Subiría la calidad media de vida. Es irrelevante que sus vidas fuesen mucho peor que nada. Esta es otra conclusión absurda.

Aunque hay otras objeciones al Principio de la Media, no las voy a formular aquí. Hemos visto lo suficiente como para darnos cuenta de que deberíamos rechazar este principio.

144. Por qué deberíamos rechazar la apelación a la desigualdad

Reconsideremos A y A +. La gente extra en A + tiene vidas que vale la pena vivir, y no afecta a nadie más. ¿Es A + peor que A?

La existencia de la gente extra baja la calidad media de vida. Pero, como he argumentado, este hecho es irrelevante. Cuando esta media se baja por Mera Adición, no podemos apelar convincentemente al Principio de la Media.

Hay otro modo en que se podría decir que A + es peor que A. En A + existe lo que llamo *desigualdad natural*. La gente extra resulta menos favorecida que el otro grupo, sin falta alguna de su parte. La gente extra no conoce este hecho, y no hay injusticia social. Esta desigualdad natural inadvertida, ¿hace a A + peor que A?

Un objetor podría apelar a principios de justicia como los que defiende Rawls. Uno de tales principios es

> *Maximin*: El mejor resultado es aquel en el que las personas más desfavorecidas son más favorecidas.

El Apéndice H muestra que Maximin puede entrar en conflicto con otros principios rawlsianos, pero aquí podemos ignorar esto. A diferencia de Rawls, algunos aplican este principio a toda clase de casos. Estas personas podrían afirmar que A es mejor que A +, porque A es el resultado en el que las personas más desfavorecidas son más favorecidas.

Supongamos, primero, que causar que se exista pueda beneficiar. Según esta asunción, si A + se realiza, esto beneficiará a las personas más desfavorecidas en A +. Puede parecer malo que en A + haya desigualdad. Pero lo que causa la desigualdad beneficia a todas las personas que son más desfavorecidas que las otras. Y este es el resultado que, de los que son posibles, beneficia a estas personas en la mayor medida. Podemos apelar aquí de manera convincente a otro principio rawlsiano. Según este principio, la desigualdad no es mala si lo que la causa da los mayores beneficios a todas las personas que son más desfavorecidas. Esto apoya la afirmación de que A + no es peor que A.

Supongamos a continuación que causar que se exista no pueda beneficiar. Asumiendo esto, si se realiza A + no se benefician las personas más desfavorecidas en A +. ¿Podría justificar una apelación a Maximin la afirmación de que A + es peor que A?

Como argumenté anteriormente, el Método Contractualista Ideal no se debería aplicar a la cuestión de cuántas personas deben existir. Lo mismo ocurre con Maximin. Supongamos que aceptamos este principio en casos en que, en los diferentes resultados, existiría el mismo número de personas. No se sigue que debiéramos extender Maximin a casos en que, en los diferentes resultados, existirían diferentes números de personas. Lo que implica Maximin en estas dos clases de casos es muy diferente.

En un Caso del Mismo Número, con dos resultados, consideremos el resultado en que

(1) El grupo más desfavorecido resulta más favorecido.

Si comparamos este resultado con el otro, tiene que ser verdadero

(2) que en este resultado hay *más* personas que resultan más favorecidas que el grupo más desfavorecido del otro resultado.

En Casos del Mismo Número, (2) es el *único* modo en que, en uno de dos resultados, (1) puede ser verdadero. Si (2) fuese falso,

el grupo más desfavorecido sería en los dos resultados igualmente grande, y saldría igualmente mal parado. De acuerdo con Maximin, el mejor de dos resultados es aquel en que (1) es verdadero. En Casos del Mismo Número, (1) es verdadero si y sólo si (2) es verdadero. Por eso podemos afirmar que, según Maximin, el mejor de dos resultados es aquel en que (2) es verdadero.

Ahora comparemos A y A +. Si apelamos a Maximin en este Caso de Diferente Número, este implica que A es mejor que A +, puesto que A es el resultado en el que (1) es verdadero. Cuando sólo hay un grupo en un resultado, este grupo es tanto el más favorecido como el más desfavorecido. ¿Es (2) verdadero en el resultado A? ¿Hay más personas en A que resultan más favorecidas que el grupo más desfavorecido en A +? No las hay. En Casos de Diferente Número, hay *otro* modo en que, en uno de los dos resultados, (1) puede ser verdadero. (1) puede ser verdadero porque

> (3) en este resultado no existen ciertas personas que, en el otro resultado, llevan vidas que vale la pena vivir.

(3) y (2) son diferentes. Como (1) es verdadero de este modo diferente, no podemos asumir simplemente que, en Casos de Diferente Número, una apelación a Maximin esté justificada. Es una cuestión abierta si, como (1) es verdadero de este modo diferente, esto hace mejor el resultado. Si en uno de dos resultados (2) es verdadero, esto es claramente un buen rasgo de este resultado. Es sin duda bueno que haya más personas que resulten más favorecidas que el grupo más desfavorecido del otro resultado. Si en uno de dos resultados (3) es verdadero, ¿es esto claramente un buen rasgo? ¿Es sin duda bueno que no existan ciertas personas que, en el otro resultado, llevarían vidas que vale la pena vivir? No puede decirse que esto sea sin duda bueno. Pienso que, si (3) es verdadero, esto *no* es un buen rasgo. La verdad de (3) no hace mejor este resultado. Puesto que (3) es verdadero, (1) también lo es. Resulta más favorecido el grupo más desfavorecido. Pero, si *así* es como (1) es verdadero, (1) no es un buen rasgo. Es moralmente irrelevante que el grupo más desfavorecido resulte más favorecido de *este* modo. En

esta clase de Casos de Diferente Número no deberíamos apelar a Maximin.

Puede que algunos cuestionen estas afirmaciones. Pero, si apelan a Maximin en casos de esta clase, se enfrentarán a objeciones como las que se dirigen contra el Principio de la Media. Consideremos una variante del caso en el que Sólo Sobrevive Francia. Supongamos que si todas las demás naciones dejaran de existir *bajaría enormemente* la calidad de vida de los franceses supervivientes. Pero los franceses aún resultarían más favorecidos que la nación más desfavorecida si ninguna nación dejara de existir. Por eso Maximin implica que sería mejor que sólo sobreviviera Francia. En esta versión de este caso, la conclusión es aun más absurda. ¿Cómo podría ser mejor que todas las demás naciones dejasen de existir, con la consecuencia de que los supervivientes resultarían *mucho más desfavorecidos*?

Puesto que implica esta conclusión absurda, no debería aplicarse Maximin a este tipo de Casos de Diferente Número. Una apelación a Maximin no puede apoyar la afirmación de que A + es peor que A.

Hay partidarios de la igualdad que no apelan a Maximin. Los mismos podrían afirmar: «Como no crees que sería mejor que viviera la gente extra, no puedes creer que haya algún modo en que A + sea mejor que A. Y deberías admitir que A + es peor de un modo. Es un mal rasgo de A + que algunas personas resulten más desfavorecidas que otras, sin falta alguna de su parte. Como A + no es de ningún modo mejor que A, y de un modo es peor, A + tiene que ser peor que A».

Esto puede parecer convincente. Pero puede ser contestado del modo en que contesté a una objeción anterior. Yo estaba suponiendo que, como la población real del mundo es ahora grande, no sería mejor que se vivieran vidas extra dignas de ser vividas. Y sugerí la siguiente objeción a esta opinión: «Comparemos la población real del mundo con una población posible mucho mayor. En esta población mayor, todo el mundo tiene una calidad de vida mucho más alta, aunque cada vida contiene un intenso sufrimiento. Siempre

que imagines que esta población mayor es aún mayor, este resultado no es de ningún modo mejor, según tu opinión. Pero hay un modo en que es peor, porque habría más sufrimiento intenso. Como este resultado no es de ningún modo mejor y sí es de un modo peor, tienes que conceder que tiene que ser peor. Y tienes que conceder que, si esta población fuera lo suficientemente grande, el aumento de sufrimiento pesaría más que todos los buenos rasgos de este resultado. Aunque todas las personas en este resultado resultarían mucho más favorecidas de lo que estamos nosotros ahora, su existencia sería peor que la de la población real del mundo. Supuesta tu manera de ver las cosas, no puedes evitar esta conclusión ridícula».

Sugerí cómo podríamos evitar esta conclusión. Estuve de acuerdo en que, si hay sufrimiento intenso en un resultado, es un mal rasgo. Pero negué que *cualquier* manera de evitar este rasgo malo haría mejor el resultado. Hay como mínimo dos modos en que podría haber más sufrimiento. Podría ser verdadero o bien (1) que las personas existentes sufriesen más, o (2) que haya gente extra viviendo, cuyas vidas, si bien dignas de vivirse, contengan algún sufrimiento. De estas dos formas en que podría haber más sufrimiento, sólo (1) hace peor el resultado. Si hay más sufrimiento porque (2) es verdadero, el hecho de que haya más sufrimiento no hace el resultado peor. No sería mejor que hubiera menos sufrimiento porque esa gente extra no existiera. Sería mejor sólo si existiera, y sufriera menos.

Se dijo arriba que la desigualdad en A + es un mal rasgo. Acepto esta afirmación. Pero una vez más niego que cualquier forma de evitar este rasgo malo haría mejor el resultado. Que la desigualdad haga peor el resultado depende de cómo se realiza. Podría ser verdadero o bien (3) que ciertas personas existentes llegaran a salir peor paradas que otras, o bien (4) que haya gente extra viviendo que, aunque sus vidas son dignas de vivirse, resulta más desfavorecida que ciertas personas existentes. Sólo (3) hace el resultado peor.

Cuando (4) es verdadero, la desigualdad puede estar producida por lo que llamo Mera Adición. Hay Mera Adición cuando hay gente extra viviendo con vidas dignas de ser vividas, que no afecta a nadie más, y cuya existencia no lleva consigo injusticia social.

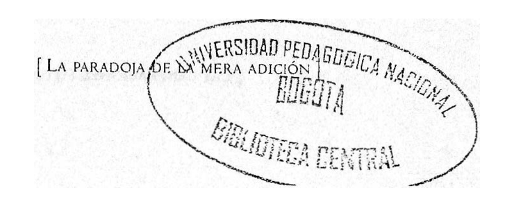
Cuando es la Mera Adición la que produce desigualdad, esta no hace peor el resultado. Esta desigualdad se evitará si o bien (5) la gente extra existe y no resulta más desfavorecida que nadie más, o bien (6) la gente extra nunca existe. Sólo (5) haría el resultado mejor. No sería mejor que no hubiera desigualdad porque la gente extra no existiera. Sería mejor sólo si la gente extra existiera, y resultara tan bien parada como cualquiera otra.

Como la desigualdad en A + está producida por Mera Adición, no hace a A + peor que A. No podemos afirmar verosímilmente que la gente extra nunca debería haber existido, *simplemente porque, sin que lo sepan, hay otras personas que resultan aún más favorecidas.*

145. La primera versión de la paradoja

Ahora comparemos A + con B. Puede ser útil hacer esta comparación a través del mundo intermedio, B Dividido, en el que las dos mitades de la población de B no pueden comunicarse. Claramente, B es tan bueno como B Dividido. Ahora podemos preguntar, «Si A + fuera a convertirse en B Dividido, ¿sería esto un cambio a mejor o a peor?».

Como los dos grupos no pueden comunicarse, el cambio no sería el resultado de una redistribución deliberada. El cambio de A + a B Dividido sería el resultado de sucesos naturales, que afectan al entorno. Y podría tener lugar lentamente, a lo largo de dos siglos. En un cambio de A + a B Dividido, la mitad más desfavorecida ganaría más de lo que perdería la mitad más favorecida. Según nuestros supuestos morales corrientes, sería un cambio a mejor. Como se trata de un Caso del Mismo Número, podemos recurrir a Maximin y al Principio de Igualdad. B Dividido es mejor que A + según ambos principios. Recordemos que el grupo más desfavorecido resulta más desfavorecido sin falta alguna de su parte. Y podemos suponer que son menos favorecidos no exactamente porque sean menos felices o tengan una calidad de vida más baja, sino porque tienen también una parte de los recursos más pequeña. Podemos suponer que B Dividido es mejor que A + tanto en tér-



minos de igualdad de bienestar como en términos de igualdad de recursos. (Asumo que, en todos mis casos imaginarios, la calidad de vida, el nivel de felicidad, y la parte de los recursos que le corresponde a cada persona se elevan y caen conjuntamente.)

Hay otra razón para juzgar que B Dividido es mejor que A +. En ciertos casos, la igualdad entra en conflicto con la beneficencia. Ocurriría así cuando un grupo más desfavorecido gana *menos* de lo que un grupo más favorecido pierde. Pero B Dividido es mejor que A + *tanto* según cualquier principio de igualdad *como* según cualquier principio plausible de beneficencia. El grupo más desfavorecido gana *más* de lo que el grupo más favorecido pierde.

Podría decirse que los principios de igualdad son de aplicación únicamente en el interior de una sociedad, donde puede haber injusticia social. Si esto es cierto, B Dividido es aún mejor que A + según cualquier principio de beneficencia. Y esta afirmación sobre la igualdad no es plausible. Consideremos

> *Rico y Pobre.* Supongamos que sé de dos personas que viven en diferentes sociedades que no pueden comunicarse. La persona a la que llamo *Rico* se halla en mejor situación que la que llamo *Pobre*. Como no vaya a ayudar a Rico, disminuirá su calidad de vida y su participación en los recursos. Yo puedo o intervenir para mantener a Rico en su nivel actual, o en vez de eso ayudar a Pobre. Si ayudo a Pobre, puedo hacer que suba al nivel en el que, sin mi ayuda, caería Rico. Y Pobre subiría más de lo que Rico caería.

La mayoría de nosotros pensará que, en caso de ayudar a alguien, sería mejor ayudar a Pobre en vez de a Rico. Y la mayoría de nosotros pensará que esto haría mejor el resultado, tanto porque Pobre ganaría más de lo que perdería Rico, como porque entonces ninguno estaría en peor situación que el otro. La mayoría de nosotros pensaría así, aunque Rico y Pobre vivieran en dos sociedades que (a no ser a través de mí) no se pueden comunicar. De modo que pensaríamos que B Dividido es mejor que A +. Como B es obviamente tan bueno como B Dividido, B también es mejor que A +.

Supongamos que pensáramos que A + no es peor que A. Ahora pensamos que B es mejor que A +. Estas creencias juntas implican

que B no es peor que A. B no puede ser peor que A si es mejor que algo —A +— que no es peor que A. Pero antes pensábamos que B *es* peor que A. Tenemos tres creencias que son inconsistentes e implican contradicciones. Implican que B es y no es peor que A. A esto lo llamo la *Paradoja de la Mera Adición* [44].

Esta no es exactamente un conflicto entre diferentes principios morales. Podemos tener una moralidad pluralista en la que pensemos que sería mejor tanto que hubiese mayor igualdad como que hubiese una suma mayor de beneficios. Puede haber entonces casos en que una mayor igualdad reduciría la suma de beneficios. Aquí nuestros dos principios sí que entrarían en conflicto. Pero no habría inconsistencia alguna en nuestra concepción moral. Simplemente tendríamos que preguntarnos si, dados los detalles del caso, la ganancia en igualdad sería más o menos importante que la pérdida de beneficios. Aquí intentaríamos alcanzar un juicio que tuviera en cuenta todos los factores. En la Paradoja de la Mera Adición, las cosas son diferentes. Aquí estamos inclinados a creer, *teniendo en cuenta todos los factores*, que B es peor que A, aunque B sea mejor que A +, que no es peor que A. No se pueden creer estas tres cosas consistentemente, desde el momento en que implican contradicciones. Una de estas creencias tiene que irse.

¿Cuál debería irse? ¿Podemos decir honestamente que creemos que habría sido mejor que nunca hubiera existido el grupo extra en A +? ¿O podemos decir honestamente que creemos que un cambio de A + a B no sería un cambio a mejor? Si afirmáramos esto último, estaríamos diciendo que lo que importa más es la calidad de vida de las personas más favorecidas. Si la calidad de vida de estas cae, esto no queda moralmente compensado ni siquiera por una ganancia mayor en la calidad de vida de un grupo más desfavorecido e igualmente grande. Esto es así aunque el grupo más desfavorecido no esté en peor situación a causa de alguna falta de su parte. Llamemos a esta la *Concepción Elitista*. Según ella, lo que les ocurre a las personas en mejor situación importa *más* que lo que les ocurre a

[44] La primera parte de este capítulo repite Parfit (7): la segunda parte presenta un argumento nuevo.

los más desfavorecidos. Una versión más radical de esta concepción sería *Maximax*. Es lo contrario de Maximin. Según esta concepción, deberíamos dar absoluta prioridad al mantenimiento o a la elevación de la calidad de vida de las personas más favorecidas. Ambas concepciones se aplican al mundo real. Pocos de nosotros las encontrarían aquí moralmente aceptables.

Puede pensarse que, si apelamos al Nivel Sin Valor o a la Concepción Léxica, ya hemos aceptado esta Concepción Elitista. Pero no es así. Recurrimos al Nivel Sin Valor cuando consideramos resultados en que no hay desigualdad. Supongamos que las vidas en A se hallan por encima del Nivel Sin Valor, y que las vidas en B y en Z están por debajo de este nivel. Podríamos afirmar entonces que la existencia de A tiene valor moral intrínseco, mientras que no habría tal valor en la existencia de B ni de Z. Por eso B y Z serían peores que A. Como no habría desigualdad si lo que se realizara fuese A, esta afirmación podría ser aceptada por los partidarios de la igualdad.

Cuando comparamos B con A +, estamos comparando resultados en uno de los cuales hay desigualdad. Según las afirmaciones que acabamos de hacer, sólo las vidas de los miembros del grupo más favorecido en A + tienen valor moral intrínseco. Aunque pensemos esto, no necesitamos afirmar que A + sea mejor que B. Nuestros Principios de Igualdad y de Beneficencia implican que B sería mejor que A +. Y nuestras declaraciones sobre el valor intrínseco podrían ser anuladas por estos otros principios. Es una concepción diferente y más elitista la de que, en un mundo en que *existe* un grupo más desfavorecido, y no lo es a causa de una falta por su parte, lo que importa más es la calidad de vida del grupo más favorecido. Esta Concepción Elitista entra en conflicto con nuestros Principios de Igualdad y Beneficencia, y los anula.

La Concepción Elitista no está implicada ni por la Concepción Léxica ni por la apelación al Nivel Sin Valor. Sin embargo, podría haber versiones elitistas de estas dos opiniones. No necesitamos ser elitistas en todos los casos. Podríamos afirmar simplemente que, si las vidas en A se hallan por encima del Nivel Sin Valor, y las vidas en B están por debajo del mismo, B sería peor que A +. Según este modo de ver las cosas, no nos opondríamos a toda redistribución

entre los que están en mejor y en peor situación. Podríamos conceder que, en la mayor parte de los casos, una pérdida para las personas más favorecidas podría ser compensada por una ganancia mayor para las personas que han resultado menos favorecidas. Nos opondríamos a tal redistribución sólo cuando causase que las personas más favorecidas cayeran por debajo del Nivel Sin Valor. Afirmaciones similares se aplican a la Concepción Léxica. Según la versión elitista de esta concepción, nos opondríamos a la redistribución sólo cuando causase que las personas más favorecidas cayesen por debajo del Nivel Maravilloso.

Supongamos que, cuando comparamos B con A +, no pudiéramos aceptar ninguna versión de la Concepción Elitista. Pensamos que B es mejor que A +. Si no podemos pensar que A + es peor que A, tenemos que concluir que B *no* es peor que A. Tenemos que concluir que, si estos fueran dos futuros posibles para una sociedad, no sería peor que el que se realizase fuera B: el doble de personas que estuvieran *todas* en peor situación.

La Paradoja de la Mera Adición no nos fuerza a aceptar esta conclusión. La podemos evitar si rechazamos una de nuestras otras dos creencias. Tal vez, aunque las encontremos difíciles de rechazar, encontramos aún más difícil de aceptar que B no es peor que A. Supongamos que decidimos que, de los dos modos de evitar esta conclusión, lo que es menos difícil de creer es que A + es peor que A. Entonces podemos mantener nuestra idea de que B es peor que A.

Deberíamos tomar nota, sin embargo, de que no podemos simplemente afirmar que A + *tiene* que ser peor que A, puesto que es peor que algo —B— que es peor que A. Estaríamos rechazando aquí una de tres afirmaciones inconsistentes simplemente sobre el fundamento de que no es consistente con las otras dos. Esto podría decirse contra *cada* afirmación. Para evitar la Paradoja, tenemos que comparar sólo A y A +, dejando de lado el resto del argumento, y tenemos que creer que A + es peor. Tenemos que creer que es malo en sí mismo que la gente extra viva alguna vez. Tenemos que creer que esto es malo, aunque estas personas tengan vidas dignas de ser vividas, y aunque no afecten a nadie más. En la medida en que

encontramos que esto es difícil de creer, todavía nos enfrentamos a una paradoja.

Puede objetarse: «Tu argumento conlleva una especie de truco. Cuando comparas A y A + afirmas que la existencia del grupo extra no va a ser peor para nadie. Pero, en el momento en que nos hemos desplazado a B, el grupo original ha pasado a estar en peor situación. La adición del grupo extra *es* peor para el grupo original. Esta es la razón por la que A + es peor que A».

El argumento puede reformularse. Supongamos que estamos considerando posibles estados del mundo de hace muchos siglos, por ejemplo en el Siglo Noveno. No hay razón para temer las consecuencias futuras; sabemos lo que ocurrió después. Supongamos a continuación que A + fuese el estado real del mundo en este siglo pretérito. Entonces podemos preguntar si habría sido mejor que el estado real hubiera sido A. Al preguntar esto, podemos suponer que A + *no* cambió después a B. La existencia del grupo más desfavorecido en A + *no* afectó para peor al grupo más favorecido. Y, como los grupos no se podrían comunicar, no había injusticia social. Dados estos hechos, ¿era peor A + de lo que habría sido A? ¿Fue malo que el grupo más desfavorecido existiera en absoluto? También podemos hacer otra pregunta. El mundo no cambió, en efecto, de A + a B. Pero, si lo hubiera hecho, ¿habría sido un cambio a mejor? Según esta versión del argumento, la última objeción queda socavada. La existencia del grupo más desfavorecido no fue peor para el grupo más favorecido. Así las cosas, puede que parezcamos obligados a admitir que A + no fue un estado de hecho peor de lo que habría sido A. Y puede que parezcamos obligados a admitir que un cambio de A + a B habría sido un cambio a mejor. De estas dos afirmaciones se sigue que B no habría sido peor que A [45].

[45] F. Myrna Kamm y J. L. Mackie parecen indicar ambos que, mientras que podría ser nuestro deber, sobre fundamentos igualitaristas, cambiar A + en B, este cambio no sería una mejora. Puede que tengamos el deber de hacer lo que produzca el resultado peor. Esta concepción aportaría una solución parcial a la Paradoja de la Mera Adición. Pero no sería una solución completa.

Hay otra objeción a este argumento. Algunos dicen: «Tenemos que distinguir dos casos. Si no fuera posible para A + cambiar a B, A + no sería peor que A. Si este cambio *fuera* posible, A + *sería* peor que A» [46].

En esta última versión del argumento, podemos añadir la asunción de que el cambio de A + a B habría sido posible. Si el estado real fuera A +, sería difícil de creer que A habría sido mejor. Y sería difícil de negar que, si un cambio de A + a B hubiera sido posible, esto habría representado una mejora. Según la objeción que acabamos de dar, si el cambio de A + a B hubiese sido posible, aunque no ocurriera, deberíamos cambiar nuestra opinión sobre A y A +. Si un cambio que no ocurrió pudiera haber ocurrido, *sí* que habría sido mejor que la gente extra nunca hubiera existido.

Si estuviéramos discutiendo lo que la gente debe hacer, una afirmación semejante podría ser plausible. Que deba obrar de una de dos maneras puede depender de si sería posible para mí obrar de una tercera manera. Pero tales afirmaciones no son plausibles cuando las aplicásemos a resultados del tipo que estoy discutiendo. Estos no son los diferentes resultados previsibles de un conjunto de actos que son posibles para una persona o grupo de personas. Ninguna persona, o grupo, elige si el resultado real será A, A + o B. Yo supuse que A + era el estado real del mundo en un siglo pretérito, y supuse también que A + no cambió efectivamente a B. Entonces pregunté si, comparado con A +, A habría sido mejor. La bondad relativa de estos dos resultados no puede depender de si podría haber ocurrido un tercer resultado que nunca ocurrirá.

Ahora se puede decir: «Supongamos que estos resultados *fueron* los efectos previsibles de diferentes actos posibles. Si preguntamos lo que debemos hacer, resolvemos la Paradoja. Asumamos que podríamos ocasionar A, o A + o B. Sería incorrecto ocasionar A +. Sería incorrecto porque hay un resultado mejor, B, que podríamos haber ocasionado. Pero también sería incorrecto ocasionar B, puesto que hay un resultado mejor: A».

[46] Esta objeción la apuntan Tooley, Woodford y otros.

Como veremos después, la Paradoja podría referirse a lo que debemos hacer. Pero hay una respuesta más sencilla a esta objeción. No resuelve nuestra Paradoja. Simplemente la ignora. Cualquier paradoja puede ignorarse. Esto no es solución.

Podemos añadir estas afirmaciones. La mayor parte de nuestro pensamiento moral puede que sea sobre lo que debemos hacer. Pero también tenemos pareceres sobre la bondad y maldad relativas de los diferentes resultados. Como he dicho, no se trata de opiniones sobre la bondad o la maldad morales, en el sentido que se aplica a los actos o a los agentes. Si un terremoto mata a miles de personas, esto no es moralmente malo en este sentido. Pero es malo en un sentido que tiene relevancia moral. Nuestras opiniones acerca de la bondad relativa de los diferentes resultados a veces dependen de nuestras opiniones acerca de lo que debemos hacer. Pero tal dependencia a menudo toma la dirección inversa. Como la última objeción misma muestra, algunas de nuestras creencias sobre lo que debemos hacer dependen de nuestras creencias sobre la bondad relativa de los resultados. Como estas últimas creencias forman la base de buena parte de nuestra moralidad, no podemos rehusarnos a considerar un argumento que trata sobre estas creencias. No podemos ignorar, como hace esta última objeción, la bondad relativa de A y A +. Por eso esta objeción no resuelve la Paradoja.

Hacemos frente a la Paradoja si creemos que la Mera Adición no puede hacer el resultado peor. Algunos piensan que la Mera Adición hace el resultado *mejor*. Afirman que A + es mejor que A. Estas personas aceptarían también mi afirmación de que B es mejor que A +. Estas dos afirmaciones implican que B es mejor que A.

Si aceptamos estas afirmaciones, y rechazamos la Concepción Elitista, no podemos evitar la Conclusión Repugnante. Hay un resultado posible *C* cuya relación con B es igual que la relación de B con A. En C hay el doble de personas, más desfavorecidas todas ellas que todo el mundo en B. El argumento de arriba puede volverse a aplicar. Si concluimos que B es mejor que A, tenemos que concluir que C es mejor que B. Según el mismo argumento, D sería mejor que C, E mejor que D, y así sucesivamente alfabeto abajo. El *mejor* resultado sería Z: una población enorme, todos cuyos integrantes tienen vidas que apenas valen la pena de vivirse.

146. Por qué todavía no estamos obligados a aceptar la Conclusión repugnante

Puede parecer que, aunque simplemente afirmamos que A + no es peor que A, estamos obligados a aceptar la Conclusión Repugnante. Puede parecer que, si B fuera mejor que A +, que no es peor que A, B tiene que ser mejor que A. Por el mismo razonamiento, C tiene que ser mejor que B, D mejor que C, y así sucesivamente.

Este razonamiento asume que *no peor que* implica *al menos tan bueno como*. Esta es una asunción natural. Pero, tras reflexionar, vemos que aquí no está justificada. Consideremos un resultado que es como A +, salvo que la gente extra tiene una calidad de vida de algún modo más alta. Llamémosle A + *Mejorado*. Claramente, este resultado es mejor que A +. Si pensamos tanto que A + no es peor que A como que A + Mejorado es mejor que A +, ¿tendremos que concluir que A + Mejorado es mejor que A? No. Podemos afirmar que, mientras que A + Mejorado es mejor que A + , los dos, simplemente, *no son peores que A* [49].

Como *no peor que* no tiene por qué implicar *al menos tan bueno como*, esta última afirmación es coherente. Y en muchas otras áreas estas son la clase de afirmaciones que debemos hacer. Consideremos tres candidatos para un premio literario, un novelista y dos poetas. Podríamos afirmar, del novelista y del primer poeta, que ninguno es peor que el otro. Esto no sería lo mismo que decir que no los podemos comparar. Sería lo mismo, antes bien, que afirmar una posibilidad de comparación aproximada. Hay muchos poetas que serían peores candidatos que este novelista, y muchos novelistas que se-

[49] Debo este punto a R. M. Dworkin y A. K. Sen. Tanto yo como muchos otros lo habíamos pasado por alto por más de diez años.

rían peores candidatos que el primer poeta. Estamos afirmando, de estos dos, que algo importante puede decirse de sus méritos respectivos. Ninguno es peor que el otro. Están al mismo nivel. Supongamos a continuación que juzgamos que el *segundo* poeta es ligeramente mejor que el primero. (Cuando comparamos a dos poetas, nuestro juicio puede ser menos aproximado.) ¿Nos obliga este juicio a concluir o bien que el segundo poeta es mejor que el novelista, o bien que el primero es peor? No. Podemos afirmar que, aunque el segundo poeta sea mejor que el primero, ninguno es peor que el novelista, que no es peor que ninguno. Podemos afirmar de mis imaginarios estados del mundo, de una manera parecida, que A + Mejorado es mejor que A +, pero ninguno es peor que A, ni tampoco es A peor que ninguno.

La posibilidad de comparación aproximada es simplemente, en algunos casos, el resultado de la ignorancia. Cuando esto es así, pensamos que en principio hay posibilidad de comparación precisa o plena. Esto sería verdadero, cuando comparamos al novelista con uno de los dos poetas, si las únicas posibilidades fuesen que uno es mejor o que los dos son exactamente igual de buenos. Lo cual no es plausible en un caso como este. La posibilidad de comparación aproximada es aquí *intrínseca*, no el resultado de la ignorancia. ¿Tiene que ser verdadero, de Proust y Keats, o que uno fue el escritor más importante, o que los dos fueron *exactamente igual* de importantes? Ni siquiera en principio podría haber una precisión tal. Pero algunos poetas son escritores más importantes que algunos novelistas, y más importantes en más o en menos. Shakespeare es un escritor muchísimo más importante que P. G. Wodehouse, pero Swinburne no es, en el mejor de los casos, muchísimo más importante. Tal posibilidad de comparación aproximada intrínseca se cumple, pienso, tanto para la bondad de ciertas clases de resultado, como para la cuestión de si una persona está, de formas moralmente significativas, en peor situación que otra [50].

Cuando hay sólo posibilidad de comparación aproximada, *no peor que* no es una relación transitiva. (Una relación *R* es transitiva

[50] Véase la discusión de la comparabilidad parcial en Sen (1).

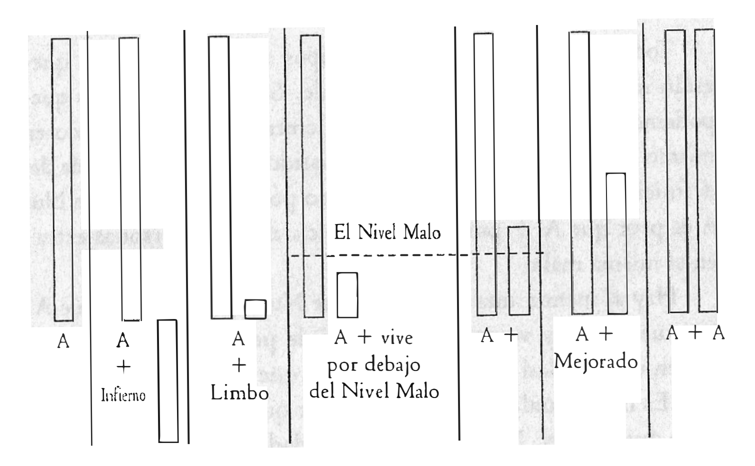
si, cuando X está R-relacionado con Y, e Y está R-relacionado con Z, X tiene que estar R-relacionado con Z.) El primer poeta no es peor que el novelista, que no es peor que el segundo poeta. Esto no nos obliga a cambiar nuestra opinión de que el primer poeta *es* peor que el segundo.

Supongamos que creemos tanto que B es mejor que A + como que A + no es peor que A. Puesto que *no peor que* no implica aquí *al menos tan bueno como*, no estamos obligados a concluir que B es mejor que A. Sólo podemos concluir que B no es peor que A. Y deberíamos comprobar que, al alcanzar esta conclusión, no hemos asumido la transitividad de *no peor que*. No lo hemos hecho. Hemos decidido que A + *no* es peor que A, pero *es* peor que B. Hemos concluido que B no puede ser peor que A. Esta conclusión estuvo justificada. Asumamos lo contrario. Asumamos que B es peor que A. Como A + es peor que B, y B es peor que A, A + tiene que ser peor que A. Este argumento es válido puesto que, a diferencia de *no peor que*, *peor que* es transitiva. Pero rechazamos la conclusión de este argumento. Pensamos que A + *no* es peor que A. Como también creemos que A + *es* peor que B, tenemos que rechazar la otra premisa de esta conclusión. Tenemos que concluir que B no puede ser peor que A.

Se nos puede forzar a admitir esta conclusión. Pero no se nos puede obligar a ir desde aquí ni siquiera a una forma debilitada de la Conclusión Repugnante. Es cierto que, por el mismo razonamiento, C no puede ser peor que B, D no puede ser peor que C, y así sucesivamente. Pero como *no peor que* no es transitiva, podemos afirmar que, mientras que C no es peor que B, que no es peor que A, C *es* peor que A.

147. LA APELACIÓN AL NIVEL MALO

Hay un argumento mejor a favor de la Conclusión Repugnante. Antes de que lo formule, discutiré otra concepción. Consideremos la gama de resultados que se muestra a continuación.



En A + Infierno el grupo extra tiene vidas (libres de pecado) que son mucho peor que nada. Si pudieran suicidarse se suicidarían. Evidentemente, A + Infierno *es* peor que A, en el sentido moralmente relevante. Y es también evidente en la misma medida que A + A *no* es peor que A. En algún lugar entre los dos tenemos que cambiar de idea. ¿Dónde debería venir el cambio? Algunos dirían que en el nivel donde las vidas de la gente extra se hacen dignas de ser vividas. Según esta opinión, A + Limbo no es peor que A.

Podríamos en vez de esto aceptar una idea que sugirió Kavka. Él llama a ciertas clases de vida *reducidas*, y afirma que, si no intervienen otros factores, es «intrínsecamente indeseable desde un punto de vista moral» que las tales vidas sean vividas [51]. Si alguien vive una vida reducida, habría sido mejor en sí mismo que esta persona nunca hubiera vivido, y que nadie hubiera existido en su lugar. (Como incluye la expresión «en sí mismo», esta afirmación no se extiende a los efectos sobre otras personas.)

La plausibilidad de la idea de Kavka depende de lo que cuente como reducido. Kavka llama a una vida reducida cuando es «significativamente deficiente en uno o más de los aspectos más importantes que generalmente hacen a las vidas humanas valiosas y dignas de ser vividas». Añade que una vida así «será por regla general digna de

[51] En Kavka (4).

ser vivida, consideradas las cosas en su conjunto». ¿Sería plausible esta concepción cuando la aplicásemos a vidas que son *perfectamente* dignas de ser vividas?

Consideremos el hecho de tener descendencia, uno de los «aspectos más importantes que generalmente hacen las vidas humanas... dignas de ser vividas». Hay quienes, a pesar de su esterilidad, tienen una vida perfectamente digna de ser vivida. Dejando a un lado los efectos en otros, ¿es malo que tales personas vivan en absoluto? No. Consideremos a continuación una discapacidad severa que dura toda la vida. Algunos ciegos llevan una vida perfectamente digna de ser vivida. Dejando a un lado los efectos en los otros, ¿es malo que tales personas vivan en absoluto? ¿Habría sido mejor que nunca hubieran vivido, y que nadie hubiera vivido en su lugar? Una vez más, la respuesta es No. Consideremos a continuación una discapacidad cuyos efectos sean más graves. Supongamos que, ya que una persona la tiene, *no* es verdad que su vida sea *perfectamente* digna de ser vivida, o que esté siquiera próxima a ser perfectamente digna de ser vivida. La opinión de Kavka es aquí más plausible. Puede haber personas cuyas vidas, aunque dignas de vivirse, se hallen tan minadas por la enfermedad o la privación que, incluso aparte de los efectos en otros, sea malo que vivan en absoluto.

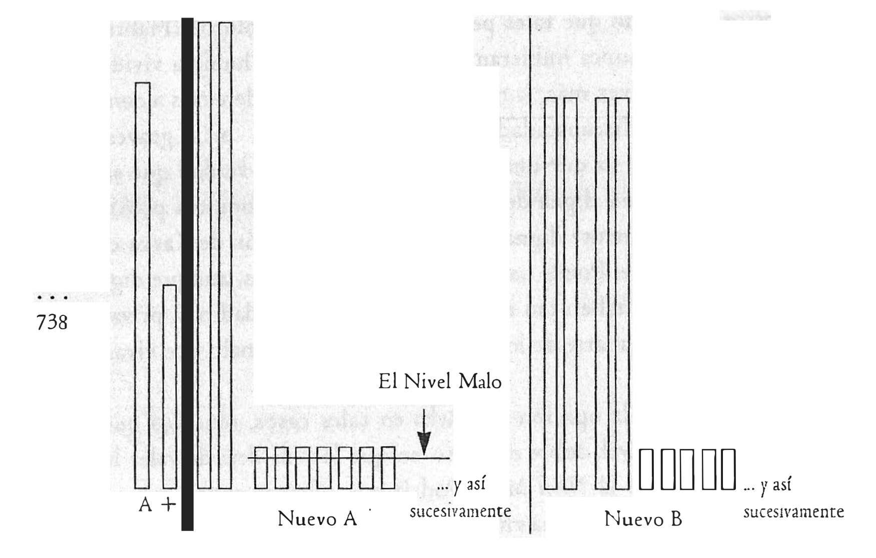
Si aceptamos la opinión de Kavka en tales casos, tenemos que introducir otro nivel, sobre el punto en que la vida deja de valer la pena. Llamémosle el *Nivel Malo*. Podríamos afirmar ahora que es malo que cualquier vida sea vivida en este nivel o por debajo de este nivel. Aunque tal vida sea digna de ser vivida, y sea de valor para la persona cuya vida es, habría sido en sí mismo mejor que nunca se hubiera vivido.

Para algunos de nosotros la concepción de Kavka puede proporcionar una respuesta parcial a la Paradoja de la Mera Adición. Podemos pensar que el Nivel Malo está por encima del nivel donde la vida deja de valer la pena. Entonces eludiremos la Paradoja en los casos en que el grupo extra en A + vive vidas que no están por encima de este Nivel Malo. Pero, aunque esto consiga algo, no consigue mucho. La concepción de Kavka no es plausible cuando se aplica a

vidas que se hallan siquiera cerca de ser perfectamente dignas de ser vividas. Las vidas malas tienen que ser peor que esto. Tienen que ser gravemente deficientes en todos los rasgos que pueden hacer a una vida digna de ser vivida. Aunque dignas de vivirse, tienen que estar llenas de obstáculos y ser de baja calidad.

148. La segunda versión de la paradoja

Consideremos los resultados que se muestran debajo



Cada bloque representa diez mil millones de personas. De manera que A + contiene veinte mil millones de personas. En esta versión de A +, hasta el grupo más desfavorecido tiene una calidad de vida *extremadamente alta*.

En Nuevo A existen muchísimos grupos extra de personas. Supongamos que estos grupos viven en planetas de otros sistemas solares. Nuevo A es un resultado en el futuro distante. Aunque estos grupos vinieron todos de la Tierra, ahora no pueden comunicarse fácilmente.

Todas las personas en estos grupos extra tienen vidas que no están muy por encima del Nivel Malo. Sus vidas son tales que no podemos afirmar honestamente que creemos que habría sido en sí mismo mejor que nunca hubieran existido. Esto se desprende de mi definición del Nivel Malo. Por eso no podemos pensar que Nuevo A es peor que A + porque la existencia de estas personas extra sea en sí misma mala.

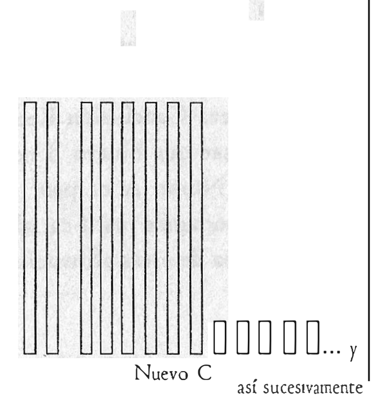
Hay al menos una manera en que Nuevo A es mejor que A +. En nuevo A hay veinte mil millones de personas, todas las cuales tienen una calidad de vida más elevada que nadie en A +.

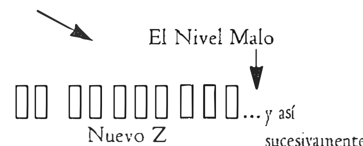
¿Es la desigualdad en Nuevo A peor que la desigualdad en A +? Creo que es *mejor*. No hay más desigualdad entre los dos grupos más favorecidos. Y la desigualdad restante está producida por Mera Adición. Como sostuve, cuando se produce de esta forma, la desigualdad no hace peor el resultado. Puesto que esta desigualdad no hace peor el resultado, y ya no hay desigualdad entre los grupos más favorecidos, Nuevo A es mejor que A + en términos igualitarios. No hay otro rasgo que pudiera decirse que haga a Nuevo A peor que A +. Como Nuevo A es de dos maneras mejor que A +, y de ninguna manera peor, Nuevo A es *mejor* que A +.

Mis afirmaciones sobre la desigualdad pueden negarse. Según algunas opiniones, la desigualdad en Nuevo A es peor que la desigualdad en A +. Pero no puede decirse que sea mucho peor. Y, aunque haya una manera en que Nuevo A es peor que A +, hay otra manera en que es mejor: el hecho de que las personas más favorecidas tienen una calidad de vida más alta. Cuando comparamos estos dos rasgos, no podemos afirmar convincentemente que Nuevo A sea peor que A +. Si negamos que Nuevo A es mejor, al menos tenemos que admitir que no es peor que A + [53].

Comparemos ahora Nuevo A con Nuevo B. Esto es como la comparación entre A + y B, salvo por los grupos adicionales que no son afectados. Puesto que existen estos grupos adicionales, no es aquí verdadero que, si Nuevo A cambiara a Nuevo B, esto aboliría

[53] A menudo no está claro si determinado cambio hace mejor o peor cierta desigualdad. Esta cuestión está bien discutida en Temkin, y es probable que sea publicada.





la desigualdad natural. Pero sería verdadero que, aunque el grupo más favorecido perdería, un grupo más desfavorecido ganaría *varias veces más en la misma medida*. Esta ganancia relativa es mucho mayor que en el argumento anterior.

Como antes, este cambio no sería el resultado de una redistribución deliberada. Se realizaría de alguna manera natural, tal vez a causa de cambios en el entorno. A no ser que aceptemos Maximax, o alguna versión de la Concepción Elitista, no podemos afirmar que este cambio haga el resultado peor. Una ganancia muchísimo mayor de la gente que está en peor situación tiene que contar más que una pérdida muchísimo menor de la gente más favorecida. A no ser que seamos elitistas, tenemos que juzgar por consiguiente que Nuevo B es mejor que Nuevo A. En casos en que todavía hay alguna desigualdad, existen diversas opiniones sobre qué modelos de desigualdad serían mejores o peores. Aunque estas opiniones no coincidan en muchos casos, todas las que son plausibles estarían de acuerdo en que la desigualdad en Nuevo B es menos mala que la desigualdad en Nuevo A. Y, como antes, Nuevo B no es sólo mejor que Nuevo A en términos igualitarios. Nuevo B es mejor según cualquier principio plausible de beneficencia. Si hubiera un cambio de Nuevo A a Nuevo B, los grupos más desfavorecidos ganarían *muchísimo más* de lo que perderían los grupos más favorecidos.

Ahora comparemos Nuevo B con Nuevo C. Una vez más, los grupos más favorecidos resultarían más desfavorecidos. Pero estos grupos resultarían más desfavorecidos en una cantidad muchísimo más pequeña que aquella en la cual el mismo número de grupos más desfavorecidos resultaría más favorecido.

Por el mismo razonamiento, Nuevo C es mejor que Nuevo B. Ese razonamiento nos lleva a Nuevo Z. En este resultado hay una población enorme, las vidas de cuyos miembros no se hallan muy por encima del Nivel Malo. Nuevo Z tiene que ser mejor que Nuevo A, puesto que cada paso Nuevo Alfabeto abajo se ha considerado un cambio a mejor, y *mejor que* es transitiva.

Recordemos que, como argumenté, Nuevo A es mejor que A +. Tomadas juntas, estas afirmaciones implican

> *La Nueva Conclusión Repugnante*: En el primero de dos resultados posibles habría dos grupos de diez mil millones de personas. Un grupo tendría una calidad de vida *mucho* más elevada que la de cualquier vida real que se haya vivido. Aunque tenga una parte más grande de los recursos, este grupo resulta, inevitablemente, más desfavorecido que el otro. El otro tiene una calidad de vida que es *aún más elevada*. En el segundo resultado posible, habría un número enorme de personas, cuya calidad de vida no está muy por encima del Nivel Malo. De estos dos resultados, el segundo sería mejor.

Algunos pueden pensar que Nuevo A no es mejor que A +, sino que simplemente no es peor. Estas personas tienen que aceptar una versión debilitada de esta nueva conclusión. Tienen que afirmar que, de A + y Nuevo Z, el segundo no sería peor.

Esta nueva conclusión es en un aspecto menos repugnante que la Conclusión Repugnante. En Z las vidas de las personas apenas eran dignas de vivirse. En Nuevo Z, las vidas de las personas son de algún modo mejores. Pero esta nueva conclusión me parece muy repugnante. Las vidas que no están muy por encima del Nivel Malo no pueden ser perfectamente dignas de ser vividas, o no pueden estar próximas a ser perfectamente dignas de ser vividas. Aunque dignas de ser vividas, tienen que estar privadas de la mayor parte de lo que hace a la vida digna de vivirse. Si no podemos evitar esta nueva conclusión, esto socava lo que la mayoría de nosotros piensa

cuando consideramos la superpoblación. Creeríamos que, si hubiera veinte mil millones de personas, todas con una calidad de vida muy alta, esto sería un resultado mejor que si hubiera en cambio muchas más personas, todas con vidas que, aunque dignas de vivirse, estuvieran llenas de privaciones y llenas de obstáculos, y fuesen de baja calidad —no muy por encima del nivel en que sería en sí mismo malo que la vida fuese vivida—. Según la Nueva Conclusión, el primero de estos resultados sería *peor*. Según la versión debilitada de esta conclusión, el primer resultado *no* sería mejor.

¿Podemos resistir este nuevo argumento? Podría sugerirse que, aunque Nuevo B fuese mejor que Nuevo A, y Nuevo C fuese mejor que Nuevo B, este razonamiento no se aplicaría todo el camino abajo del Nuevo Alfabeto. Podría decirse: «Cuando todos los grupos resultan más desfavorecidos, el Principio de Igualdad tiene un peso menor. Si dos grupos están ambos en buena situación, una ganancia mayor para el grupo más desfavorecido pesa más moralmente que una pérdida más pequeña para el grupo más favorecido. Pero el Principio de Igualdad tiene un peso menor cuando se aplica a grupos que se hallan en peor situación. Y hay un nivel por debajo del cual este principio no tiene peso».

Esta concepción carece de plausibilidad. Mucha gente piensa que el Principio de Igualdad tiene un peso diferente cuando se aplica a grupos que están todos en peor situación. Pero estas personas consideran que, en estos casos, el principio tiene un peso *mayor*. Consideran lo contrario de la concepción que se ha sugerido. Y no sé de nadie que la acepte [54].

Una vez más, deberíamos recordar que el argumento no apela sólo a la igualdad. En cada uno de los pasos Nuevo Alfabeto abajo, los grupos más favorecidos perderían *muchísimo menos* de lo que ganaría otro número igual de grupos más desfavorecidos. Esto sería un cambio para mejor tanto según el Principio de Igualdad como según cualquier principio plausible de beneficencia. A no ser que seamos elitistas, tenemos que admitir que cada cambio es a mejor.

[54] Véase de nuevo Temkin.

Si consideramos que Nuevo B sería mejor que Nuevo A, no podemos negar de un modo convincente afirmaciones comparables sobre resultados contiguos en lugares más bajos del Alfabeto. Si queremos eludir ambas versiones de la Nueva Conclusión Repugnante, tenemos por tanto que o bien afirmar que Nuevo B no es mejor que Nuevo A, o bien afirmar que Nuevo A es peor que A +. Como he argumentado, a no ser que podamos justificar alguna versión de la Concepción Elitista, ninguna de estas afirmaciones es defendible.

Ahora resumiré el argumento. Los grupos extra en Nuevo A viven vidas que se hallan por encima del Nivel Malo. Dada mi definición de este nivel, no podemos pensar que sea en sí mismo malo que tales vidas se vivan. La existencia de estas personas no afecta a nadie para peor. Y la existencia de estas personas no introduce desigualdad natural. He argumentado que, con respecto a la igualdad, Nuevo A es mejor que A +. Los que no están de acuerdo no pueden decir que, en este aspecto, Nuevo A es mucho peor. Y, en otro aspecto, Nuevo A es mejor que A +. Veinte mil millones de personas tienen una calidad de vida más alta que la de las personas más favorecidas en A +. Dados estos hechos, no puede afirmarse que Nuevo A sea peor que A +. Tampoco podemos apelar a la afirmación de que, comparado con A +, Nuevo A puede tener peores consecuencias. Podríamos decir que, si los grupos en Nuevo A llegan a ser más capaces de comunicarse, habrá injusticia social, que debe ser eliminada, y que el resultado de la redistribución sería peor que A +. He explicado cómo se puede bloquear esta afirmación. Podemos suponer, como rasgo de nuestro caso, que tal redistribución nunca ocurrirá. El resultado real será, y seguirá siendo, Nuevo A. Nunca habrá un cambio ni a Nuevo B ni a Nuevo Z. Estos otros resultados seguirán siendo meramente posibles. Como esto es así, tenemos que comparar simplemente el valor relativo de Nuevo A y A +. ¿Sería mejor que la gente extra no viviese nunca, a costa de una calidad de vida *más baja* para *todas* las personas que *van a* vivir? Esto es difícil de creer. Y es difícil tanto creer que Nuevo B no habría sido mejor que Nuevo A, como tener tal creencia acerca de dos resulta-

dos contiguos en un lugar más bajo de la serie. Si tenemos una creencia así, tenemos que aceptar alguna versión de la Concepción Elitista. Si la aceptamos, tenemos que aplicarla al mundo real. Supongamos que pensamos que la vida de algunos europeos no se halla muy por encima del Nivel Maravilloso. Según la versión elitista de la Concepción Léxica, si una pérdida para estas personas las desplazara por debajo del Nivel Maravilloso, esta pérdida no sería moralmente compensada ni siquiera por una ganancia muchísimo mayor de las personas que están en una situación mucho peor —como por ejemplo los niños africanos que sufren desnutrición. Esto es difícil de creer. Pero también es difícil de creer que Nuevo Z es mejor que A +. Nuestro problema persiste.

149. La tercera versión

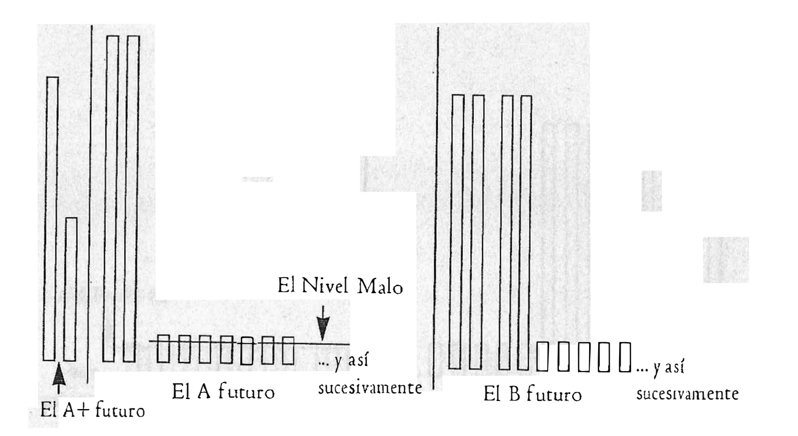
Necesitamos considerar casos que contengan, en los diferentes resultados, a todas las personas que alguna vez vivan. El último diagrama puede mostrar un caso así: puede ser como el que se muestra abajo.

Este diagrama muestra diferentes futuros posibles. Cada bloque representa ahora mil años de lo que queda de la historia humana. La altura de cada bloque muestra la calidad de vida que todo el mundo disfruta durante todos estos mil años. En todo momento de todos estos futuros posibles habrá diez mil millones de personas viviendo.

El A + futuro comienza en el Siglo 23. Los dos siglos anteriores han ido bien. Cuando el A + futuro comienza, ya no hay más desigualdad entre las diferentes personas, y la calidad de vida es extremadamente alta. Sucesivas generaciones disfrutan de esta calidad de vida durante mil años. Entonces el sol se hace mucho más caliente. Y eso determina que la calidad de vida sea mucho más baja, en muchas maneras. Aun así, sigue siendo extremadamente alta durante un segundo milenio. Entonces el sol se hace muchísimo más caliente, acabando así con la historia humana.

En el A futuro, los primeros dos mil años van aun mejor. Todos tienen una calidad de vida más alta que la de las personas más favo-

recidas en el A + futuro. El sol no cambia durante este período. Llegando al final del mismo, los científicos predicen que el sol se hará mucho más caliente. Como consecuencia del pronóstico, la gente se pone a excavar muchas cavernas profundas. Lo cual permite a la humanidad sobrevivir al tremendo calor de la superficie de la Tierra. La gente vive en las cavernas durante muchos miles de años. La vida subterránea vale la pena, pero es mucho peor de lo que hubiera sido en la superficie terrestre. Durante los años vividos en las cuevas la calidad de vida no está muy por encima del Nivel Malo. El sol entonces estalla, poniendo punto final a la historia humana.

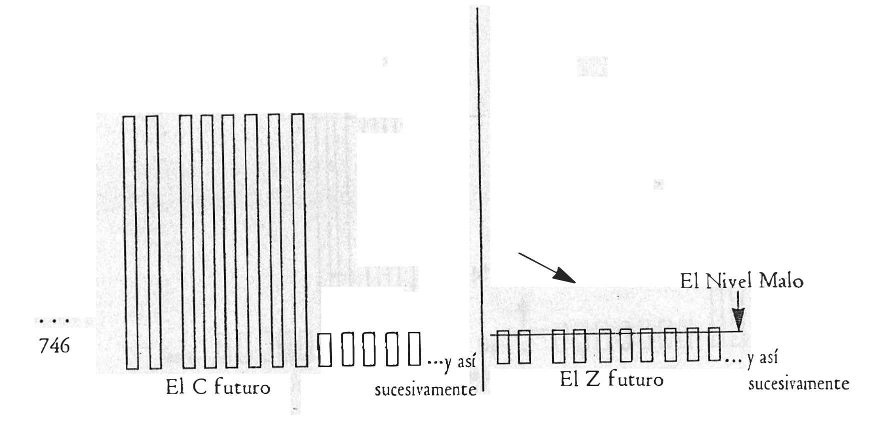


Deberíamos asumir que el A futuro es lo que realmente ocurre. Los otros futuros son simplemente posibles alternativas. Comparado con el A futuro, ¿sería mejor el A + futuro? ¿Sería mejor que no hubieran sido hechas las cavernas, de forma que la historia finalizase después de los primeros dos mil años? ¿Sería mejor esto, incluso a costa de una calidad de vida más baja durante estos dos mil años?

Las personas de las cavernas llevan una vida que es digna de ser vivida y se halla por encima de nuestro Nivel Malo. Es difícil de creer que la existencia de estas personas sea en sí misma mala. Y esto es más difícil de creer que en la versión anterior de este argumento. En el A futuro, no hay desigualdad entre las personas que viven durante algún período; hay desigualdad sólo entre las diferen-

tes generaciones. Esto refuerza la afirmación de que, como esta desigualdad está producida por Mera Adición, no hace peor el resultado.

Es difícil de creer que vaya a ser malo que sean excavadas las cavernas, de forma que exista la gente extra. Aunque esto fuese malo, no demostraría que el A futuro es peor que el A + futuro. Estos futuros difieren de otra manera. Si pensamos que el A futuro es peor, tenemos que creer que será mejor que la gente extra nunca exista, a costa de una calidad de vida *más baja* para *todos* los que *alguna vez* vivan. Esta creencia es absurda.



En esta versión del argumento, podemos preguntar qué debe hacer la gente. Consideremos a los que escapan a la muerte cavando las cavernas. Si pensamos que el A futuro es peor que el A + futuro, tal vez debiéramos concluir que estas personas no deberían tener hijos. Y tenemos que concluir que tienen una razón moral para no tener hijos, puesto que con ello harían peor el resultado. Esto no es plausible. Si estas personas tuvieran hijos, tanto estos como todos sus descendientes tendrían vidas dignas de ser vividas. Dado este hecho, es difícil de creer que tendrían *alguna* razón moral para no tener hijos. Lo cual confirma mi afirmación de que el A futuro no es peor que el A + futuro. Y, en otro aspecto, el A futuro sería mejor. En los primeros dos mil años, todo el mundo tendría una calidad de vida más alta. Como el A futuro no sería de ninguna

forma peor que el A + futuro, y sería mejor de una manera, deberíamos concluir que, de los dos, el A futuro sería mejor.

Consideremos ahora el B futuro. Este diferiría del A futuro del siguiente modo. Durante los primeros dos mil años, cuando la vida se vive sobre la superficie de la Tierra, la calidad de vida sería de algún modo inferior. Pero habría una elevación mucho mayor de la calidad de vida durante los siguientes dos mil años en las cavernas. Como indica el diagrama, aunque la vida sería diferente en la superficie y en las cavernas, la calidad de vida sería la misma para los primeros cuatro mil años.

El A futuro es lo que realmente ocurrirá. ¿Sería mejor el B futuro? Las personas más favorecidas, en los primeros dos mil años, tendrían una calidad de vida más baja. Pero perderían mucho menos de lo que ganaría un mismo número de personas en los dos mil años siguientes. Según nuestros Principios de Igualdad y de Beneficencia corrientes, esto sería un resultado mejor.

Como antes, estas dos diferencias podrían producirse a causa de cambios naturales en el entorno. Pero también podrían producirse por actos deliberados. Supongamos que el cambio en el sol fuera previsto al comienzo de los primeros dos mil años. Los que vivan en estos años podrían ser capaces, con algún coste para ellos mismos, de dar beneficios muchísimo mayores a los que viviesen en los dos mil años siguientes. Algunos dirían que, como esto haría el resultado mejor según los Principios de Igualdad y Beneficencia, es lo que estas personas deben hacer. Otros podrían decir que semejante altruismo no sería un deber, sino que sería simplemente admirable desde el punto de vista moral. Las dos opiniones apoyan mi argumento. Para oponernos a él, tendríamos que afirmar que semejante altruismo no haría mejor el resultado. Como antes, si pensamos que este resultado no sería mejor, tenemos que aceptar alguna forma de la Concepción Elitista. Tenemos que pensar que lo que les sucede a las personas más favorecidas importa más que lo que les sucede a las personas más desfavorecidas. Una pérdida para las personas más favorecidas no sería moralmente compensada por una ganancia muchísimo mayor para los que están en peor situación. La mayoría de nosotros encon-

traría esto imposible de creer. Tenemos que admitir entonces que el B futuro sería mejor que el A futuro.

Observaciones similares se aplican al C futuro. Este diferiría del B futuro de una manera similar. Las personas durante los primeros cuatro mil años saldrían perdiendo, pero habría una ganancia muchísimo mayor para las personas en los siguientes cuatro mil años. Según los Principios tanto de Igualdad como de Beneficencia, el C futuro sería mejor. Por el mismo razonamiento, también lo serían el D futuro, el E futuro, y así sucesivamente. El mejor de todos estos posibles futuros sería el Z futuro. Sería verdadero aquí que por todo el resto de la historia humana la calidad de vida no estaría muy por encima de lo que habría sido en las cavernas del A futuro. La calidad de vida siempre estaría próxima al Nivel Malo.

Como *mejor que* es una relación transitiva, el Z futuro sería mejor que el A futuro. Y, como razoné, el A Futuro sería mejor que el A + futuro. Tenemos que concluir así que el Z futuro sería mejor que el A + futuro. En comparación con un futuro en el que muchos miles de millones de personas tienen todas una calidad de vida *extremadamente alta*, sería *mejor* que, en cambio, hubiera muchas más personas, *todas* cuyas vidas no estarían muy por encima del nivel en el que pensamos que sería *malo* que estas vidas se vivieran.

En esta versión, el argumento es más fuerte. Y no eludimos las preguntas sobre lo que la gente debe hacer. Hacemos esas preguntas, y las respuestas dan apoyo al argumento. Si encontramos la conclusión de este argumento tan difícil de creer, la Paradoja es mayor [55].

[55] Para una versión revisada de este argumento, y algunas reflexiones suplementarias, véase mi «Overpopulation and the Quality of Life» [«Superpoblación y calidad de vida»], en *Practical Ethics*, ed. P. Singer, Oxford University Press, 1986.

CAPÍTULO DE CONCLUSIÓN

Cuando le preguntaron sobre su libro, Sidgwick dijo que la primera palabra del mismo era *Ética*, y la última *fracaso*. Esta podría haber sido la última palabra de mi Cuarta Parte. Como argumenté, necesitamos una nueva teoría de la beneficencia. Esta tiene que resolver el Problema de la No-Identidad, evitar las Conclusiones Repugnante y Absurda, y resolver la Paradoja de la Mera Adición. Fracasé a la hora de encontrar una teoría que pueda satisfacer estos cuatro requisitos. Aunque no logré encontrar tal teoría, creo que, si lo intentasen, otros podrían tener éxito.

En las otras partes de este libro, alcancé varias conclusiones. La mayor parte de ellas tiene un rasgo común.

150. IMPERSONALIDAD

Mis dos temas son las razones y las personas. Y he argumentado que, de diversos modos, nuestras razones para actuar deberían hacerse *más impersonales*. Una impersonalidad mayor puede parecer amenazadora. Pero muchas veces sería mejor para todos.

El capítulo 3 argumentó que, en nuestra preocupación por otras personas, la mayoría de nosotros comete errores. La mayoría de nosotros desea evitar hacer daño a otras personas. Pero muchos creen en

> (El Segundo Error.) Si un acto es correcto o incorrecto a causa de sus efectos, los únicos efectos relevantes son los de este acto particular.

Esto lleva a estas personas a ignorar lo que hacen *juntas*. Y la mayoría de nosotros cree en

> (Los Errores Cuarto y Quinto.) Si un acto tiene efectos en otros que son insignificantes o imperceptibles, no puede ser moralmente incorrecto *porque* tenga estos efectos.

Estas falsas creencias no tenían importancia en las pequeñas comunidades en que la mayor parte de la gente vivió durante la mayor parte de la historia. En estas comunidades, hacemos daño a otros sólo si hay personas a quien cada uno de nosotros daña considerablemente.

La mayoría de nosotros vive ahora en comunidades grandes. Los malos efectos de nuestros actos pueden ahora dispersarse sobre miles o incluso millones de personas. Nuestras falsas creencias son ahora errores serios. Que son errores queda claro en mi caso imaginario de los Torturadores Inofensivos. Cada uno de estos torturadores contribuye, a sabiendas pero de manera imperceptible, al dolor sufrido por cada una de mil víctimas. Estos torturadores obran muy mal. Saben que, aunque ninguno de ellos represente una diferencia perceptible, juntos infligen un gran dolor a sus víctimas.

Hay incontables casos reales de esta clase. En ellos es verdadero, del acto de cada uno, que sus efectos sobre los otros son insignificantes o imperceptibles. Y equivocadamente damos en pensar que, puesto que esto es así, los efectos de nuestros actos no pueden hacerlos incorrectos. Pero, aunque cada acto tenga efectos insignificantes, es muchas veces cierto que juntos nos causamos un gran perjuicio a nosotros mismos y a los demás. Algunos ejemplos son la

contaminación, los embotellamientos de tráfico, el agotamiento de los recursos, la inflación, el paro, la recesión, la captura abusiva de pescado, la agricultura intensiva, la erosión del suelo, la hambruna y la superpoblación.

Mientras tengamos estas falsas creencias, nuestra ignorancia es una excusa. Pero una vez que hemos visto que estas creencias son falsas, no tenemos excusa. Si seguimos actuando de estos modos, nuestros actos serán moralmente incorrectos. Algunos pueden ser tan malos como los de los Torturadores Inofensivos.

Los altruistas racionales no tienen estas falsas creencias. Si todos fuéramos altruistas racionales, sería mejor para todos. Pero los altruistas racionales son, en este sentido, más impersonales: no preguntan simplemente, «¿Será mi acto peor para alguien? ¿Se quejará alguien?». Piensan que es irrelevante que sus actos dañen perceptiblemente a cualquier otra persona.

La vida en las grandes ciudades es impersonal de manera inquietante. Y no podemos resolver este problema como no lo ataquemos en sus propios términos. Igual que necesitamos a los ladrones para coger a los ladrones, tenemos necesidad de principios impersonales para evitar los malos efectos de la impersonalidad.

El capítulo 4 argumentó que, puesto que a menudo es directa y colectivamente contraproducente, la Moralidad del Sentido Común tiene que revisarse. La versión revisada R es un paso parcial hacia el altruismo racional. Una vez más tenemos que ser más impersonales.

Consideremos nuestras obligaciones con nuestros hijos. Según la Moralidad del Sentido Común, debemos dar a nuestros propios hijos algunas clases de prioridad. De acuerdo con R, hay casos en que *no* debemos dar estas clases de prioridad a nuestros propios hijos. Debemos hacer lo que sería lo mejor para los hijos de todos, imparcialmente considerados. Al decirnos que ignoremos nuestras relaciones con nuestros propios hijos, R nos dice que ignoremos la que puede ser la más fuerte de todas nuestras relaciones personales.

Si todos seguimos este principio impersonal, y no damos prioridad a nuestros propios hijos, será mejor para todos nuestros hijos. La impersonalidad es de nuevo mejor, hasta en términos personales.

Afirmaciones similares se aplican a nuestras relaciones con personas tales como nuestros padres, nuestros amigos, vecinos, alumnos o pacientes.

La Segunda Parte argumentó que deberíamos rechazar la teoría de la racionalidad que llamamos del Propio Interés. PI es la teoría que da la mayor importancia a la diferencia entre personas, o a la *condición separada de las personas.* PI *me* dice que haga todo lo que vaya a ser lo mejor para *mí*. Para PI, las unidades fundamentales son las *diferentes vidas.* Mi preocupación suprema debería ser que toda mi vida marche tan bien como sea posible. Cada persona es racionalmente requerida a darse a sí misma, y a su propia vida, prioridad absoluta.

Como deberíamos rechazar PI, nuestra teoría tiene que ser en cierto modo más impersonal. No está obligada a afirmar que la preocupación suprema de cada persona debiera ser ella misma; y no tiene que dar importancia suprema a los límites entre vidas. Pero nuestra teoría no tiene por qué ser el Principio de la Benevolencia Imparcial de Sidgwick. Deberíamos aceptar la Teoría Crítica del fin Presente, o CP. Según ella, la unidad fundamental no es el agente durante toda su vida completa, sino el agente en el momento de actuar. Aunque CP le niega al propio interés la importancia suprema, y también se la niega a la vida completa de una persona, no es impersonal. CP establece que lo que es racional para mí hacer ahora depende de lo que quiero o valoro o creo ahora. Esta afirmación da *más* importancia a los valores o las creencias particulares de cada persona. Como CP da más importancia a lo que distingue a las diferentes personas, de este modo diferente es *más* personal que PI.

La Tercera Parte argumentó a favor de otra clase de impersonalidad. Cuando consideramos diversos casos imaginarios, descubrimos qué pensamos que somos nosotros mismos. La mayoría de nosotros cree que nuestra identidad tiene que ser siempre determinada. Pensamos que, para la pregunta «¿Estoy a punto de morir?», tiene que haber siempre una respuesta, que tiene que ser siempre, de un modo absolutamente simple, Sí o No.

Esta creencia no puede ser verdadera como no seamos entidades que existen separadamente, distintas de nuestros cerebros y cuerpos y de nuestras experiencias. La existencia continua de estas entidades tiene que ser un hecho adicional profundo, distinto de la continuidad física y psicológica, y además un hecho que o se da completamente o no se da en absoluto. Una entidad semejante sería un Ego Cartesiano. Como muestran nuestras reacciones a los casos imaginarios, la mayoría de nosotros tiene la inclinación a creer que somos entidades de este tipo. Como argumenté, esto no es así.

Puesto que no es así, no podemos explicar la unidad de la vida de una persona estableciendo que las experiencias de esta vida son todas tenidas por esta persona. Podemos explicar esta unidad únicamente si describimos las diversas relaciones que se dan entre estas diferentes experiencias, y sus relaciones con un cerebro concreto. Por tanto podríamos describir la vida de una persona de un modo impersonal, que no afirme que esta persona existe.

Según esta Concepción Reduccionista, las personas existen. Pero existen únicamente de la manera en que existen las naciones. Las personas no son *fundamentales*, como creemos equivocadamente. En este sentido, esta concepción es más impersonal.

Y esta concepción presta apoyo a determinadas afirmaciones sobre la racionalidad y la moralidad. Según las Tesis Radicales, las implicaciones serían completamente impersonales. Un autor declara que, si es verdadera la Concepción Reduccionista, nos debe ser indiferente si vivimos o morimos. Y otros dicen que no tendríamos razón alguna para estar especialmente preocupados por nuestro propio futuro, y que la mayor parte de la moralidad quedaría socavada. Estos autores piensan que únicamente el hecho adicional profundo de la identidad personal nos da razones para una preocupación especial, y únicamente él apoya la mayor parte de nuestro moralidad. No parece haber ningún argumento que refute este modo de ver las cosas. Por eso es defendible afirmar que, como ese hecho adicional no existe, no tenemos razón alguna para una preocupación especial, y la mayor parte de la moralidad carece de base.

Aunque estas Tesis Radicales son defendibles, también pueden ser negadas de forma defendible. Yo suscribo las siguientes tesis menos radicales:

> El debilitamiento de las conexiones psicológicas puede reducir tanto la responsabilidad por crímenes pasados como las obligaciones de cumplir con compromisos pasados.

> Según la Concepción Reduccionista, resulta más plausible rechazar los principios distributivos. Es más plausible centrarse no en las personas sino en las experiencias, y establecer que lo que moralmente importa es la naturaleza de las mismas. Según el Principio Utilitarista impersonal, la pregunta de *quién* tiene una experiencia es tan irrelevante como la de *cuándo* se tiene la experiencia. Este principio ignora los límites entre vidas, o la condición separada de las personas. Según la Concepción Reduccionista, este principio es más plausible. (No quiero decir «más plausible que su negación». Quiero decir «más plausible de lo que lo sería sobre la base de la Concepción No-Reduccionista». Esto resulta compatible con la afirmación de que, incluso sobre la base de la Concepción No-Reduccionista, este principio no es plausible.)

> Debemos considerar que la imprudencia grave es moralmente incorrecta. Esto reduce las pretensiones de la autonomía personal. Ya no tenemos el derecho de hacer lo que nos plazca cuando sólo resultemos afectados nosotros mismos. Es incorrecto imponernos a nosotros mismos, por ninguna razón válida, un gran perjuicio.

Estas afirmaciones, de nuevo, dan menos importancia tanto a la unidad de cada vida como a los límites entre vidas. Como antes, mis conclusiones son más impersonales.

Hay dos excepciones. Para algunos autores lo que importa en la supervivencia es la continuidad física, o sea, la existencia continua del mismo cerebro concreto. Dicen que, si yo estuviera a punto de ser teletransportado, debería considerar tal perspectiva como algo casi tan malo como la muerte. Aunque mi Réplica fuese completamente continua conmigo desde el punto de vista psicológico, no es esto lo que importa. Lo que importa es que no sería físicamente continua conmigo. No estoy de acuerdo. Creo que lo que importa es la Relación R, continuidad y/o conexividad psicológicas. Al argumentar que es esto lo que importa, y no la continuidad física, yo estaba atacando de nuevo lo que es, en un cierto sentido, una concepción más impersonal. Según ella, lo que importa es un rasgo que compartimos tanto con los simples animales como con los simples objetos físicos. Para mí, lo que importa es *lo que nos hace personas*.

Ahora añadiré una declaración similar. Para la Concepción No-Reduccionista, la unidad profunda de cada vida se halla asegurada automáticamente, por muy al azar, miope y pasivamente que se viva esta vida. Para la Concepción Reduccionista, la unidad de nuestras vidas es cuestión de grado, algo sobre lo que podemos influir. Puede que queramos que nuestra vida tenga una unidad mayor, en el sentido en que un artista puede querer crear una obra unificada. Y podemos *darle* a nuestra vida mayor unidad, de ciertos modos que expresen o realicen nuestros valores y nuestras creencias particulares. Como la Concepción Reduccionista le da mayor importancia a cómo elegimos vivir, y a qué distingue a las diferentes personas, este es un segundo sentido en que es *más* personal.

En la Cuarta Parte mis conclusiones son impersonales, en el sentido más claro y contundente posible. Si queremos evitar la Conclusión Repugnante, no podremos resolver el Problema de la No-Identidad recurriendo a un principio de *personas afectadas*. Tendremos que recurrir a uno que verse sobre la calidad y la cantidad de las vidas que se viven, pero que no verse sobre lo que es bueno o malo para las personas a las que afectan nuestros actos.

También sostuve que, si recurrimos a un principio semejante, no supone ninguna diferencia moral lo que está implicado por los principios de personas afectadas. Cuando tomé en consideración los Dos Programas Médicos, llegué a la conclusión de que era irrelevante que anular sólo uno de estos programas fuese peor para los niños afectados. Si tenemos que recurrir a un principio semejante, debiendo ignorar entonces principios de personas afectadas, esto tiene amplias implicaciones teóricas, de tipo impersonal.

La *Ética* pregunta qué resultados serían buenos o malos, y qué actos correctos o incorrectos. La *Metaética* se pregunta por el significado del lenguaje moral o por la naturaleza del razonamiento moral. También se pregunta si la Ética puede ser objetiva —si puede hacer afirmaciones que sean *verdaderas*.

Algunos dan por hecho que hay sólo dos maneras de hacer Ética, o de razonar sobre la moralidad. Una es el *Camino Bajo*, que simplemente apela a nuestras intuiciones. Otra es el *Camino Alto*, la Metaética. Si podemos ofrecer la mejor explicación de la naturaleza del razonamiento moral, podremos tener la esperanza de que esto vaya a implicar determinadas afirmaciones sobre la moralidad. Podremos esperar que nuestra Metaética implicará conclusiones en Ética.

Creo que estas no son las únicas maneras en que podemos argumentar sobre moralidad. No he tomado el Camino Alto, excepto cuando asumí que una teoría moral aceptable no puede ser directa y colectivamente contraproducente. A menudo he tomado el Camino Bajo, apelando a nuestras intuiciones. Pero uno de mis objetivos principales ha sido explorar una variedad de diferentes clases de argumentos, que se encuentran entre los extremos Bajo y Alto.

Los capítulos 1 y 4 discuten

(*a*) el argumento de que una teoría es contraproducente.

Con esta clase de argumentos, podemos hacer un progreso innegable. Como PI y C pueden ser indirectamente contraproducentes, las dos tienen que hacer nuevas afirmaciones sobre nuestros deseos y disposiciones. Como la Moralidad del Sentido Común es a menudo directa y colectivamente contraproducente, esto implica, según casi toda teoría metaética, que esta moralidad tiene que revisarse.

El capítulo 3 apela a

(*b*) hechos cuya significación moral se ha pasado por alto.

Un hecho de estos es el efecto combinado de conjuntos de actos, o de lo que nosotros hacemos juntos. Otro es el efecto combinado de lo que, individualmente, son efectos imperceptibles. Mis imaginarios Torturadores Inofensivos obran muy mal, a causa de los efectos de sus actos, aunque ninguno de ellos haga que el dolor de ninguna víctima sea perceptiblemente peor. Esto refuta el parecer de que un acto no podría ser incorrecto, a causa de sus efectos en los demás, si las personas no pudiesen notar alguna vez diferencia alguna.

El capítulo 16 recurre a la misma clase de argumentos. Recurre al hecho de que fácilmente podemos afectar a las identidades de las personas futuras. Las implicaciones de este hecho —lo que llamé el Problema de la No-Identidad— tienen una significación clara para nuestras teorías morales. Pero, con escasas excepciones, las hemos pasado por alto.

Es improbable que estos sean los únicos ejemplos de esta clase de argumentos. Puede que haya muchos argumentos de esta clase no descubiertos —muchos otros hechos cuya clara significación racional o moral ha sido sencillamente pasada por alto. Esta es otra clase de argumentos con la que podemos hacer un progreso innegable.

El capítulo 6 apela a

(*c*) una descripción más completa de lo que asume e implica una teoría

Mi primer argumento contra la teoría del Propio Interés fue poco más que una pregunta. Pero para llegar hasta ella tuve que presentar la teoría Crítica del fin Presente. Mi objetivo era aislar PI, de forma que pudiese ser juzgada por sus propios méritos. Tuve la necesidad, por tanto, de contrastar PI con CP. No fue suficiente apelar a las teorías instrumentales o deliberativas. Ninguna de estas teorías corrientes cuestiona PI de un modo que un teórico del Propio Interés no pueda ignorar. CP nos proporciona tal cuestionamiento. Y nos pone en condiciones de ver lo que PI asume de

manera implícita. Podemos preguntar qué versión de CP coincidiría con PI. Esta versión afirma que una predisposición temporalmente neutral a favor de uno mismo es supremamente racional. Afirma que esta predisposición tiene que ser nuestra preocupación dominante aun si ni la tenemos ni la queremos tener aunque conozcamos los hechos y pensemos con claridad. Cuando vemos más claramente lo que asume PI, deja de ser plausible.

El capítulo 7 apela a

(*d*) una debilidad en la estructura de una teoría.

PI es una teoría *híbrida*, puesto que requiere neutralidad temporal pero rechaza los requisitos de neutralidad entre diferentes personas. Esto no la hace incoherente. Pero es un fallo estructural que hace a PI vulnerable cuando es atacada desde las dos direcciones [1].

El capítulo 8 menciona otra clase de argumentos. Esta apela a

(*e*) las implicaciones de una conclusión metafísica —una conclusión sobre los rasgos más fundamentales de la realidad, o del universo.

Hay filósofos y físicos que aseguran que el paso del tiempo es una ilusión. Yo afirmo que la mayoría de nosotros tiene falsas creencias sobre nuestra propia naturaleza, y la naturaleza de nuestra existencia continua a través del tiempo. En caso de que podamos demostrar que tenemos tales creencias falsas, una apelación a la verdad podría prestar apoyo a determinadas afirmaciones sobre la racionalidad y la moralidad. De esta forma, como asevero, la Concepción Reduccionista proporciona otro argumento contra la teoría del Propio Interés. Y esta concepción da apoyo a diversas tesis morales.

[1] Williams presenta un argumento semejante contra el Utilitarismo de Actos, afirmando que hay un «punto de ruptura» en la estructura de esta teoría. Véase Williams (1), p. 114.

La Cuarta Parte recurre a argumentos de la clase (*e*). El Utilitarismo Clásico implica la Conclusión Repugnante, y el Principio de la Media implica conclusiones absurdas. La Cuarta Parte también apela a un argumento de la clase (*d*). Este refuta la idea de que hay un límite superior para el valor de la cantidad durante un periodo dado. Cuando esta idea se amplía para incluir el sufrimiento no compensado, aparece un fallo estructural. Como PI, es una concepción híbrida. Aunque afirma que el valor positivo de la cantidad tiene un límite superior, no puede poner límite de una manera convincente al valor negativo de la cantidad. Por eso implica otra conclusión absurda.

152. ¿Deberíamos alegrarnos de mis conclusiones o por el contrario lamentarlas?

Mantengo que

(1) Como a menudo son indirectamente contraproducentes, la teoría del Propio Interés y el Consecuencialismo tienen que hacer afirmaciones sobre nuestros deseos y disposiciones. Tienen que afirmar que deberíamos estar dispuestos a obrar de maneras que ellas declaran irracionales y moralmente incorrectas.

(2) Como la Moralidad del Sentido Común es a menudo directa y colectivamente contraproducente, tiene que ser revisada.

Como sugiero en el capítulo 5, estas dos conclusiones reducen el desacuerdo entre la Moralidad del Sentido Común y el Consecuencialismo. Este es un buen resultado. Señala hacia una teoría unificada, que eliminaría el desacuerdo.

Mantengo que

(3) Al considerar cómo afectan nuestros actos a los demás, la mayoría de nosotros comete graves errores. Por eso deberíamos cambiar muchas de las formas de actuar que tenemos ahora.

Esta es otra conclusión grata. Cuando comprendamos que son errores, obraremos más probablemente de un modo que será mejor para todos.

Mantengo que

(4) En lo que respecta a la racionalidad, deberíamos rechazar la teoría del Propio Interés, y aceptar la teoría Crítica del fin Presente. Según esta, hay deseos que son irracionales, mientras que otros pueden ser racionalmente requeridos. Supongamos que conozco los hechos, pienso con claridad, y el conjunto de mis deseos no es irracional. Sería entonces irracional para mí obrar a favor de mis propios mejores intereses si con ello fuera a frustrar lo que, en ese momento, valoro o quiero más.

(5) Puesto que deberíamos rechazar la teoría del Propio Interés, deberíamos afirmar que la imprudencia grave es moralmente incorrecta.

Estas conclusiones son más difíciles de evaluar. Considero que, en principio, (4) es otra conclusión a la que debemos dar la bienvenida. Hay al menos dos clases de actos que buscan el propio interés:

i(i) Algunos de estos actos benefician al agente, pero imponen cargas mayores a otras personas. A juicio de la teoría del Propio Interés, es irracional *no* actuar de esta manera. Si la gente dejase de pensar así, serían menos los que actuasen de esta manera, lo que haría mejor el resultado.

(ii) Algunos actos encaminados al propio interés benefician mucho al agente, sin ser peores para los demás. Sería malo que menos personas actuaran de esta manera. No conseguir obrar así es una gran imprudencia. Esto siempre es triste, y a menudo trágico. Argumenté que deberíamos ampliar nuestra teoría moral, para que declare que la imprudencia grave es moralmente incorrecta. Si aceptamos tanto (4) como (5), tal vez esto no incremente, y a lo mejor reduzca, la incidencia de la imprudencia grave. Pero puede haber personas que acepten (4) pero rechacen (5). Si la gente dejara de creer que la imprudencia grave es irracional, y siguiera creyendo que no puede ser moralmente incorrecta, esto podría tener malos efectos. Y hay personas para las que la acusación de «irracional» tiene más gravedad que la de «inmoral».

El rechazo de la teoría del Propio Interés merece la bienvenida desde otro punto de vista. Comparada con CP, PI es una verdadera amenaza para la moralidad. Hay muchos casos en que PI entra en conflicto con la moralidad. Algunos de estos conflictos son inevitables, sea lo que sea lo que queramos o valoremos. Si creemos en PI, pensaremos que, en estos casos, sería irracional obrar moralmente. Y esta creencia puede hacernos menos inclinados a obrar moralmente.

No se aplican a CP afirmaciones parecidas. Para algunas versiones de CP, hay muchos casos en que CP entra en conflicto con la moralidad. Pero este conflicto no es inevitable. Si nos preocupáramos lo suficiente por la moralidad, el conflicto desaparecería.

Mantengo que

(6) La mayoría de nosotros debería cambiar de opinión sobre la naturaleza de las personas y de la identidad personal a través del tiempo. La verdad es aquí muy diferente de lo que piensa la mayoría de nosotros.

(7) Dado este cambio en nuestras creencias sobre nosotros mismos, deberíamos cambiar también algunas de nuestras concepciones morales. Y ciertas Tesis Radicales, aunque se pueden negar con justificación, también son defendibles.

Estas conclusiones también son difíciles de evaluar. Si aceptamos las Tesis Radicales, tal vez no les demos la bienvenida a las conclusiones. Swinburne escribe que, si aceptara la Concepción Reduccionista, no tendría ninguna razón para seguir viviendo. Otros autores afirman que la mayor parte de la moralidad se vería socavada.

Yo rechazo estas Tesis Radicales. Creo que la Relación R —continuidad y conexividad— nos da una razón para estar especialmente preocupados por nuestro propio futuro. Tal vez no sea esta una razón tan fuerte como la que nos proporcionaría el Hecho Adicio-

nal. Y, puesto que la conexividad psicológica es una cuestión de grado, deberíamos rechazar la afirmación de que tiene que ser irracional cuidarse menos de algunas partes de nuestro futuro. Deberíamos rechazar la Teoría Clásica del Propio Interés. Ya he explicado por qué doy la bienvenida a esta conclusión. Si nos hacemos reduccionistas, este cambio de manera de pensar también apoya ciertos cambios en nuestras concepciones morales. Pero no encuentro preocupantes estos cambios.

Cuando me pongo a considerar qué implica (6), me alegro de que sea verdadero. Este cambio de la manera de ver las cosas también tiene sus efectos psicológicos. Me hace preocuparme menos por mi propio futuro y por el hecho de que voy a morir. Por contra, ahora me importan más las vidas de los otros. A estos efectos les doy la bienvenida. La Metafísica *puede* producir los consuelos de la filosofía.

Por último, mantengo que

(8) Ya que podemos afectar fácilmente a las identidades de las personas futuras, nos enfrentamos al Problema de la No-Identidad. Para resolverlo necesitamos una nueva teoría de la beneficencia. Teoría que tiene también que evitar las Conclusiones Repugnante y Absurda, además de resolver la Paradoja de la Mera Adición.

Como aún no he encontrado esta teoría, son poco gratas estas conclusiones. Socavan nuestras creencias relativas a nuestras obligaciones con las generaciones futuras. La mayoría de nosotros asume que la elección de una de dos políticas sociales podría ir en contra de los intereses de los que vivan en el futuro lejano. En muchos casos esta creencia es falsa. Tiene que haber una objeción moral a nuestra elección de la Política Arriesgada o de la Reducción. Pero esta objeción no puede apelar a nuestro principio corriente de la incorrección de perjudicar a los demás. Aunque estas dos políticas tengan lo que son claramente malos efectos, elegirlas no será peor para nadie.

Como fracasé en encontrar el principio al que deberíamos apelar, no puedo explicar la objeción a nuestra elección de tales políti-

cas. Soy del parecer de que, aunque hasta el momento haya fracasado, yo mismo o bien otros podríamos encontrar el principio que necesitamos. Pero hasta que esto suceda (8) es una conclusión inquietante.

Mientras tanto, deberíamos esconder este problema a los que vayan a decidir si incrementamos nuestro uso de la energía nuclear. Estas personas saben que la Política Arriesgada podría causar catástrofes en el futuro lejano. Será mejor que piensen algo falso: que la elección de la Política Arriesgada iría contra los intereses de las personas muertas por tal catástrofe. Teniendo esta falsa creencia, llegarán a alcanzar con mayor probabilidad la decisión correcta.

Hay otros sentidos en que (8) es desagradable. La mayor parte de nosotros creería que las Conclusiones Repugnante y Absurda son lo que las he llamado. Hasta que sepamos cómo evitarlas, y cómo resolver tanto el Problema de la No-Identidad como la Paradoja de la Mera Adición, tendremos creencias que no podemos justificar, y que sabemos que son inconsistentes.

Si yo u otros resolvemos estos problemas, (8) será bien recibida en un sentido trivial. Nos encanta resolver problemas. Pero antes de que hayamos encontrado soluciones, debemos lamentar esta conclusión. Con más problemas irresueltos, estamos más lejos de la Teoría Unificada, estamos más lejos de la teoría que resuelve nuestros desacuerdos, y que, puesto que logra este objetivo, podría merecer que la llamemos la verdad.

153. Escepticismo moral

Los escépticos morales niegan que una teoría moral pueda ser verdadera. Más en general, niegan que alguna teoría pudiera ser *objetivamente* la mejor teoría. Un argumento a favor de esta concepción es que, a diferencia de las Matemáticas, la Ética no es un tema en el que todos estemos de acuerdo. Puede negarse que este sea un buen argumento. Pero para minarlo tenemos que dar con una teoría que resuelva nuestros desacuerdos. Antes de que demos con ella, podemos dar otras dos razones para poner el duda el Escepticismo

Moral. Son razones para afirmar que la cuestión de la objetividad no está fijada, sino que sigue abierta.

Muchos son escépticos morales, pero no escépticos en lo referente a la racionalidad. Nos podemos dedicar mejor a la cuestión de la objetividad si consideramos no sólo razones morales sino todos los tipos de razones para actuar. Hay algunas afirmaciones que todos nosotros aceptamos.

Supongamos que, como no me mueva, me matará una roca que se me viene encima, y supongamos que lo que más deseo es sobrevivir. ¿Tengo una razón para moverme? Es innegable que la tengo. Esta afirmación habría sido aceptada en todas las civilizaciones, en todas las épocas. Esta afirmación es *verdadera*.

Como hay algunas afirmaciones verdaderas sobre razones para actuar, podemos negar lo que algunos escépticos mantienen. A veces se dice que, a diferencia de las rocas y de las estrellas, no puede *haber* valores morales objetivos. Tales entidades no pueden existir. Son demasiado extrañas como para ser parte de «la fábrica del universo». Pero, en el caso recién descrito, tengo una razón para moverme. Tal vez no sea una razón moral. Pero, como existe esta razón, puede *haber* razones. Las razones para actuar pueden, en el único sentido relevante, «existir». Como hay algunas razones para actuar, es una cuestión abierta la de si algunas de estas son razones morales [2].

Hay otra razón para poner en duda el Escepticismo Moral. No deberíamos asumir que la objetividad de la Ética tiene que ser todo-o-nada. Puede haber una parte de la moralidad que sea objetiva. Al describir esta parte, nuestras afirmaciones pueden ser verdaderas. Cuando consideramos esta parte de la moralidad, o estas cuestiones morales, podemos encontrar la Teoría Unificada que acabaría con nuestros desacuerdos. Puede haber otras cuestiones acerca de las que nunca nos pongamos de acuerdo. Puede que no haya respuestas verdaderas para las mismas. Como la objetividad no tiene por qué ser todo-o-nada, los escépticos morales pueden tener razón en

[2] En Sidgwick (1), pp. 37-8, se sugiere que podemos desafiar al escéptico recurriendo a razones no morales.

parte. *Estas* cuestiones pueden ser subjetivas. Pero esto no tiene por qué poner en duda la Teoría Unificada [3].

154. CÓMO TANTO LA HISTORIA HUMANA COMO LA HISTORIA DE LA ÉTICA PUEDEN ESTAR SÓLO EMPEZANDO

Hay quien cree que no puede haber progreso en Ética porque en este terreno ya se ha dicho todo. Al igual que Rawls y Nagel [4], yo pienso lo contrario. ¿Cuántas personas han hecho de la Ética No-Religiosa la tarea de su vida? Antes del pasado reciente, muy pocas. En la mayoría de las civilizaciones, la mayoría de las personas ha creído en la existencia de un Dios, o de varios dioses. Una extensa minoría era de hecho atea, por mucho que hiciese creer lo contrario. Pero, antes del pasado reciente, pocos ateos hicieron de la Ética la tarea de su vida. Tal vez Buda se encuentre entre estos pocos, y unos cuantos griegos y romanos de la Antigüedad. Después de más de mil años, hubo unos pocos más entre los Siglos Dieciséis y Veinte. Hume fue un ateo que hizo de la Ética parte de la obra de su vida. Sidgwick fue otro. Después de Sidgwick hubo varios ateos que fueron filósofos morales profesionales. Pero la mayoría de ellos no hicieron Ética. Hicieron Metaética. No se preguntaron qué resultados serían buenos o malos, o qué actos serían correctos o incorrectos. Se preguntaron sólo por el significado del lenguaje moral y por la cuestión de la objetividad, y escribieron sobre ello. La Ética No Religiosa ha sido estudiada sistemáticamente por muchas personas sólo a partir de los años sesenta del Siglo Veinte. Comparada con las demás ciencias, la Ética No Religiosa es la más joven y la menos avanzada.

Creo que si destruimos a la humanidad, cosa que ahora podemos hacer, este resultado será *mucho* peor de lo que la mayoría de la gente piensa. Comparemos tres resultados:

[3] Sigo a Nagel (3), pp. 97-126, y Nagel (4), caps. 9 y 14.

[4] Sigo a J. Rawls, «The Independence of Moral Theory» [«La independencia de la teoría moral»], *Proceedings and Addresses of the American Philosophical Association* (1974-5), pp. 5-22; y a Nagel, cap. 9.

(1) La paz.
(2) Una guerra nuclear que mata al 99% de la población mundial.
(3) Una guerra nuclear que mata al 100%.

(2) sería peor que (1), y (3) sería peor que (2). ¿Cuál es la mayor de estas dos diferencias? Los más piensan que la diferencia mayor es la que se da entre (1) y (2). Yo considero que la diferencia entre (2) y (3) es *muchísimo* mayor.

Mi opinión la mantienen dos grupos de personas muy diferentes. Ambos apelarían al mismo hecho. La Tierra permanecerá inhabitable durante al menos otros mil millones de años. La civilización dio comienzo sólo hace unos pocos miles de años. Si no destruimos a la humanidad, estos pocos miles de años pueden ser sólo una pequeña fracción de toda la historia humana civilizada. La diferencia entre (2) y (3) puede ser de este modo la diferencia entre esta pequeña fracción y todo el resto de esta historia. Si comparamos esta posible historia con un día, lo que ha ocurrido hasta ahora es sólo una fracción de segundo.

Uno de los grupos que comparten mi opinión son los utilitaristas clásicos. Ellos afirmarían, como hizo Sidgwick, que la destrucción de la humanidad sería con mucho el mayor de todos los crímenes concebibles. La maldad de este crimen radicaría en la vasta reducción de la suma posible de felicidad.

El otro grupo estaría de acuerdo, pero por razones muy diferentes. Los que lo integran consideran que hay poco valor en la mera suma de felicidad. Para ellos, lo que importa son los que Sidgwick denominó «bienes ideales» —las ciencias, las artes, el progreso moral, o el avance continuo hacia una comunidad mundial totalmente justa. La destrucción de la humanidad imposibilitaría logros ulteriores de estas tres especies. Esto sería radicalmente malo porque lo que importa por encima de todo serían los logros *más elevados* en estos órdenes, y estos logros más elevados vendrían en los siglos futuros.

No cabe duda de que podría haber logros más elevados en la lucha por una comunidad mundial totalmente justa. Y podrían darse logros más altos en todas las artes y las ciencias. Pero podría ser mayor aun el progreso en la que ahora es la menos adelantada de

estas artes y ciencias. Esta, he afirmado, es la Ética No Religiosa. La no creencia en Dios, admitida abiertamente por una mayoría, es un acontecimiento reciente, que todavía no se ha completado. Como es tan reciente, la Ética No Religiosa está en una fase muy temprana. Todavía no podemos predecir si, como en Matemáticas, llegaremos a estar todos de acuerdo. Puesto que no podemos saber cómo se desarrollará la Ética, no es irracional tener grandes esperanzas.

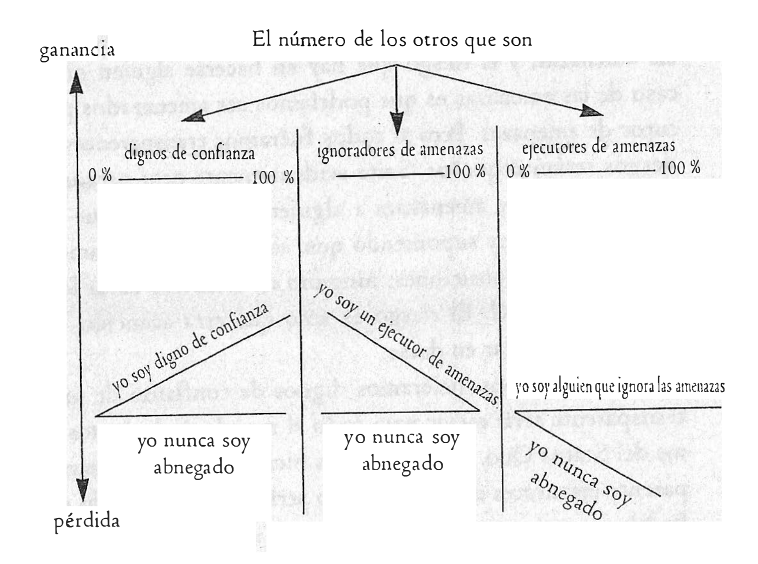
APÉNDICES

A. Un mundo sin engaño

Supongamos que todos fuéramos transparentes y no fuéramos nunca abnegados. Llamemos a esto el *Status Quo*. Probablemente sería mejor para cada uno de nosotros que se volviera digno de confianza, un ejecutor de amenazas y alguien que no hace caso de amenazas. Cada cambio implicaría ciertos riesgos, pero probablemente iban a ser compensados en gran medida por los beneficios.

¿Cómo ganaría alguien si cambiase sus disposiciones de estos tres modos? Depende de lo que hagan los demás. La ganancia derivada de volverse digno de confianza depende de cuántos de los otros se volvieran dignos de confianza. La ganancia derivada de volverse un ejecutor de amenazas depende de cuántos de los otros se volvieran personas que no hacen caso de las amenazas, y viceversa. Si todos los demás siguen sin ser nunca abnegados, no supondría ninguna ventaja ser simplemente alguien que ignora las amenazas, y como mucho una pequeña ventaja ser digno de confianza, pero sería una gran ventaja ser un ejecutor de amenazas. Alguien que es digno de confianza gana poco si nadie más es

digno de confianza [1], y gana mucho si todos los demás son dignos de confianza. Una persona que simplemente ignora las amenazas no gana nada si no hay nadie más que sea ejecutor de amenazas, y gana mucho si todo el mundo es un ejecutor de amenazas. Pero un ejecutor de amenazas gana mucho si no hay nadie que sea uno que ignora las amenazas. Estos hechos se ponen de manifiesto en el siguiente diagrama.



Se supone que todo el mundo es transparente, y que nunca es abnegado salvo en la medida en que adquiere alguna de estas tres disposiciones. El diagrama ignora los riesgos, y otras ciertas complicaciones. Podríamos evitarlas haciendo asunciones adicionales. Pero no las necesitamos aquí, porque no afectarían al argumento. Alguien que no haga caso de las amenazas podría ser un caso especial de un ejecutor de amenazas, uno que ha amenazado con ignorar las amenazas de las demás personas. Llamemos a alguien *simplemente* una persona que ignora las amenazas si esta es la única amenaza que ejecutaría.

[1] B. Hooker corrigió mi anterior opinión de que, si alguien fuese digno de confianza, no ganaría nada si nadie más fuese digno de confianza. Alguien que no fuese digno de confianza podría darme un beneficio porque confía en que yo le diese algún beneficio a cambio.

Como muestra el diagrama, si alguien gana por el hecho de volverse o digno de confianza o un ejecutor de amenazas, estas pueden ser ganancias con respecto al Status Quo. Tal persona se vuelve más favorecida de lo que lo habría quedado si ella y todos los demás hubieran seguido sin ser nunca abnegados. Pero la ganancia que se obtiene del hecho de volverse simplemente alguien que ignora las amenazas no puede elevar a nadie por encima del Status Quo. Sólo puede evitarle hundirse más.

Estos hechos pueden explicarse de la siguiente manera. Cuando alguien gana por ser digno de confianza, con frecuencia será cierto que los demás también ganan. Y estas ganancias no tienen por qué implicar pérdidas para los demás. Pueden resultar de mantener acuerdos mutuamente ventajosos, que crean nuevos beneficios sin coste alguno para los demás. Esto ocurriría, por ejemplo, con algunas formas cooperativas de la industria o de la agricultura. Pero cuando alguien obtiene ganancias del hecho de ser un ejecutor de amenazas, esto es peor para alguien más. La ganancia del ejecutor de amenazas puede ser sólo una ganancia defensiva, que evite que un aspirante a agresor obtenga una ganancia de una agresión. Pero esto sería peor para este agresor. Y cuando alguien sale ganando por ser una persona que ignora las amenazas, esto es sólo la evitación de una pérdida. Si soy de modo transparente alguien que ignora las amenazas, los ejecutores de amenazas no pueden ganar nada amenazándome. Será peor para ellos si me amenazan, porque yo ignoraré sus amenazas y ellos las ejecutarán, lo cual será peor para todos nosotros. Como sería peor para ellos que me amenazaran, no lo van a hacer. Pero si simplemente soy alguien que ignora las amenazas, mis únicas ganancias son de esta clase —lo que no pierdo para los ejecutores de amenazas. Por eso esta disposición no puede elevarme sobre el Status Quo.

Estos hechos tienen las implicaciones siguientes. Que todo el mundo se hiciera digno de confianza sería mejor para todos que si nadie se hiciera. Pero no habría semejante ganancia general si todos nosotros llegáramos a ser personas que ignoran las amenazas y que

además las ejecutan. Esto explica por qué, de estas tres desviaciones de la disposición de no ser nunca abnegado, sólo la primera de ellas se piensa que está requerida por la moralidad. Es una afirmación plausible la de que, si podemos afectar a nuestras disposiciones, moralmente debemos determinarnos a nosotros mismos a ser o a seguir siendo dignos de confianza. Pero no podría afirmarse convincentemente que, si ahora nadie es nunca abnegado con respecto a las amenazas, moralmente debemos determinarnos a nosotros mismos a ser personas que ignoran las amenazas y además las ejecutan.

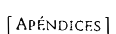
Aunque esta afirmación general no sea plausible, podríamos hacer otras dos. Si hay otras personas que tienen intenciones lo suficientemente malas, quizás debamos convertirnos en transparentes ejecutores de amenazas, para que podamos disuadirlas. (Si no somos transparentes, sería mejor desde el punto de vista moral simplemente aparentar ser ejecutores de amenazas. Esto está clarísimo en el caso de la disuasión nuclear.) También podríamos decir que, si hay otras personas que se han vuelto ejecutoras de amenazas, y tienen malas intenciones, moralmente debemos determinarnos a nosotros mismos a ser personas transparentes que ignoran las amenazas.

En el mundo tal y como es ahora, donde somos en parte opacos, sería difícil convencer a los demás de que realmente *somos* personas que no hacen caso de las amenazas y que además las ejecutan. No bastaría con ejecutar o ignorar alguna amenaza, a un coste pequeño para nosotros mismos. Como sería mejor para nosotros aparentar ser de los que ignoran las amenazas y además las ejecutan, puede ser racional para nosotros en términos del propio interés pagar este pequeño coste, en un intento de ganar esta apariencia útil. Pero este mismo hecho haría a la gente dudar de que ejecutaríamos o pasaríamos por alto las amenazas cuando supusiera un gran coste para nosotros. Un ejecutor de amenazas, por tanto, debería saludar el desarrollo de pruebas infalibles de detección de mentiras.

¿Cuáles son los riesgos que entrañan estos tres cambios en nuestras disposiciones? Si no somos transparentes, un riesgo que hay en hacerse digno de confianza es que podríamos ser engañados para mantener un acuerdo mutuo por los que meramente parecen

ser dignos de confianza pero que no van a cumplir con su parte. Si todos somos transparentes, existe sólo el riesgo más pequeño de que aquellos con quienes hacemos tales acuerdos, aunque intenten hacer su parte, puedan ser de hecho incapaces de hacerla. Este riesgo sería compensado en gran medida por los beneficios probables de la fiabilidad, esos que crea el mantenimiento de los acuerdos mutuamente ventajosos. El riesgo que hay en hacerse un ejecutor de amenazas es que podríamos amenazar a alguien que no hace caso de las amenazas, y el riesgo que hay en hacerse alguien que no hace caso de las amenazas es que podríamos ser amenazados por un ejecutor de amenazas. Pero si todos fuéramos transparentes estos dos riesgos serían pequeños. Sería evidentemente peor para un ejecutor de amenazas que amenazara a alguien transparente que ignora las amenazas. Y estoy suponiendo que, salvo cuando actuamos a partir de estas tres disposiciones, ninguno de nosotros haría lo que cree sería peor para él. El riesgo es sólo que esta asunción ocasionalmente pueda fallar en darse.

Que todos nos hiciéramos dignos de confianza de una manera transparente sería mejor para todo el mundo. Subiríamos por encima del Status Quo. Que todos nos hiciéramos de una manera transparente ejecutores de amenazas no sería mejor para todo el mundo. Podría ser mejor para algunas personas, aquellas que son débiles por naturaleza. Si ninguno de nosotros es nunca abnegado, los fuertes podrán explotar a los débiles no con amenazas sino con advertencias. Que los fuertes dañen a los débiles porque los débiles no han hecho concesiones puede que no sea peor para los fuertes. Pero si los débiles tuvieran armas que pudieran destruirles tanto a sí mismos como a los fuertes, podría ser mejor para ellos hacerse transparentes ejecutores de amenazas. Haciendo amenazas defensivas creíbles podrían salvarse a sí mismos de la explotación por parte de los fuertes. Esta ganancia para los débiles puede ser una de las razones por las que el general Gallois dio la bienvenida a la proliferación nuclear. Pero esta ganancia sería insegura: quedaría abolida si los fuertes se volvieran personas transparentes que ignoran las amenazas. Y si todos nosotros nos hiciéramos personas que ignoran las amenazas y además las ejecutan, habría mayores riesgos para todo el





mundo. Esto podría hacer a estos cambios en nuestras disposiciones peores para todos nosotros.

Esta última afirmación puede que parezca entrar en conflicto con otra de las mías. Afirmé que sería mejor para cada uno de nosotros determinarse a sí mismo a convertirse en alguien que ignora las amenazas y que además las ejecuta. Si estos cambios fuesen peores para todos nosotros, puede parecer que esto no podría ser verdadero.

Esto no es así. Estas afirmaciones podrían ser verdaderas ambas. Podría ser verdadero tanto (1) que, hagan lo que hagan los demás, sería mejor para cada uno de nosotros que se convirtiese en una persona que ignora las amenazas y además las ejecuta, como (2) que, si todos nosotros más bien que ninguno hiciéramos estos cambios, sería peor para todos nosotros. Lo que cada uno ganaría haciendo estos cambios puede ser menos que lo que perdería si todos los demás hicieran lo mismo.

Supongamos que estamos en el Status Quo, no siendo nadie nunca abnegado, y no siéndolo de una manera transparente. Si añadimos una asunción, no sería mejor para nadie que se convirtiera a sí mismo en una persona que ignora las amenazas y que además las ejecuta. Esta asunción es que, si alguien se cambiara a sí mismo de estas dos maneras, con ello ocasionaría que todos los demás hicieran los mismos dos cambios. Supongamos que esto no fuese verdadero. Sería poco probable si fuéramos muy numerosos —miembros de una sociedad con una población grande.

Si nadie más me copiara, sería mejor para mí convertirme de un modo transparente en una persona que no hace caso de las amenazas y que además las ejecuta. Si todos los demás siguen sin ser nunca abnegados, yo no gano nada siendo una persona que ignora las amenazas, pero gano mucho siendo un ejecutor de amenazas. Así es como podría elevarme en el mayor grado por encima del Status Quo. A medida que los demás comienzan a adquirir estas dos disposiciones, yo gano menos siendo un ejecutor de amenazas, pero comienzo a ganar siendo una persona que las ignora. Cuando todos los demás hayan efectuado ambos cambios, yo dejaría de estar por encima del Status Quo. Como hay ciertos

riesgos, aún puedo hundirme más. Y ahora no salgo ganando nada siendo un ejecutor de amenazas. Como todos los demás están ahora dispuestos a ignorar mis amenazas, estas se han hecho inútiles. Y como existe el riesgo de que yo podría, estúpidamente, hacer una amenaza, ahora puede ser mejor para mí perder esta disposición. Pero ahora gano mucho siendo una persona que no hace caso de las amenazas. Sería muy malo para mí no ser nunca abnegado en un mundo de ejecutores de amenazas. Todos los demás podrían entonces explotarme haciendo amenazas. Quizás la mayoría de la gente, siendo de buen natural, no lo haría. Pero unos pocos sí. De modo que sería evidentemente mejor para mí seguir siendo de una manera transparente una persona que no hace caso de las amenazas.

Mis conclusiones son, entonces, estas. En un mundo sin engaño, sería probablemente mejor para cada uno de nosotros dejar de no ser nunca abnegados, al menos de dos modos. Sería muy probablemente mejor para cada uno si se convirtiera en, y siguiera siendo, una persona digna de confianza que no hace caso de las amenazas. De acuerdo con PI, sería racional para cada uno de nosotros cambiarse a sí mismo de estos dos modos.

Como he dicho, podría ser verdadero que no podamos llevarnos a nosotros mismos a actuar de formas que creamos irracionales. Y podría ser verdadero que los demás no creyesen que nosotros actuaríamos de formas que creemos irracionales. Si una de estas fuese verdadera, PI nos diría que sería racional para nosotros cambiar no sólo nuestras disposiciones sino también nuestras creencias sobre la racionalidad. Cada uno de nosotros debería aún creer que es por lo general irracional para él hacer lo que piensa que será peor para sí mismo. Pero debería tratar de resolverse a creer que tales actos son racionales cuando implican ignorar amenazas, o cumplir promesas.

[Que a través de la tecnología podríamos hacernos transparentes, y que deberíamos reflexionar de antemano sobre tales cambios, lo aprendí de J. Glover. Para pensamientos más profundos en estas direcciones, ver Glover (3).]

B. CÓMO MI CONCLUSIÓN MÁS DÉBIL DERROTARÍA EN LA PRÁCTICA A PI

Según la más débil de mis dos conclusiones, cuando PI y P entran en conflicto, sería racional seguir cualquiera de las dos. Esta conclusión sería sumamente perjudicial para la teoría del Propio Interés. Formalmente, el resultado sería simétrico. PI habría perdido su derecho a ser la única teoría verdadera—o la mejor teoría. Habría perdido la afirmación más atrevida

(PI12) Es irracional actuar a sabiendas en contra del propio interés de uno mismo simplemente para lograr lo que, en el momento de obrar y después de una deliberación ideal, uno quiere o valora más.

Y se quedaría con

(PI13) Es racional actuar a sabiendas a favor del propio interés de uno mismo, aunque uno sepa que esto va a frustrar lo que, después de una deliberación ideal, uno quiere o valora más.

La teoría del fin Presente no habría ganado la afirmación más atrevida

(PI1) Es irracional actuar a favor del propio interés de uno cuando uno sabe que esto frustrará lo que, tras una deliberación ideal, uno quiere o valora más.

Habría ganado sólo

(PI2) Es racional hacer lo que uno sabe que va a lograr mejor lo que, tras una deliberación ideal, uno quiere o valora más, aunque uno sepa que esto va en contra de su propio interés.

Pero, aunque haya simetría formal, las afirmaciones comparables tienen para estas dos teorías diferente importancia. Es más importante para PI hacer la más contundente de sus afirmaciones. Al perder (PI13) ha perdido lo que necesita en su conflicto con P. No basta con haber mantenido el más débil (PI12).

Antes de defender esta afirmación, consideraré una similar relativa al conflicto entre PI y la moralidad. Sidgwick creía que, si preguntamos qué tenemos más razones de hacer, el resultado es formalmente simétrico. La mejor teoría moral no puede derrotar a, ni ser derrotada por, la teoría del Propio Interés sobre la racionalidad. Cuando la moralidad entra en conflicto con el propio interés, sería racional seguir cualquiera de los dos [2].

[2] G. Harman pone en duda que esta sea una concepción defendible. Podría decirse: «Tanto la moralidad como el propio interés aportan razones para actuar. Cuando la moralidad entra en conflicto con el propio interés, puede haber respuesta a la pregunta de qué es racional hacer. Una razón muy fuerte guiada por el propio interés tiene un peso mayor que una razón moral débil, y viceversa. Sólo en determinados casos ninguna razón será más fuerte que la otra». Según esta concepción, estas dos clases de razón siempre pueden ser pesadas en una balanza neutral. Y las respuestas difieren en los diferentes casos. Hay *conmensurabilidad*, con *pesos que difieren*.

Supongamos que aceptamos alguna versión de la Moralidad del Sentido Común. Entonces podríamos contestar: «Estas afirmaciones son incorrectas. Cualquier teoría moral plausible toma en cuenta el propio interés del agente. Supongamos que prometo proporcionarte un beneficio verdaderamente banal. Entonces sucede que, para mantener esta promesa, tengo que sufrir una gran pérdida. No es este un caso en que una débil razón moral sea superada por una fuerte razón interesada. En este caso, cuando llega a ser verdadero que yo sólo podría cumplir esta promesa a este grandísimo coste, deja de serlo que yo deba cumplirla. En un caso así, la moralidad *no* entra en conflicto con el propio interés. Parecidas observaciones se aplican a todos los principios morales. Que debamos actuar sobre la base de estos depende en parte de cuánto tendríamos que sacrificar. Lo que de otro modo sería nuestro deber moral deja de serlo cuando exige de nosotros un sacrificio demasiado grande, o una interferencia demasiado grande en nuestras vidas. Ya que la moralidad le da la importancia debida a las demandas del propio interés, no puedes afirmar que una débil razón moral sería superada por una fuerte razón interesada. Si debes obrar de cierto modo, aunque el coste para ti sea elevado, a ese coste se le ha dado su importancia debida. A pesar de este coste, esto es lo que debes hacer». Según este modo de pensar, las razones del propio interés se pesan en la balanza moral. Algunos de los que adoptan esta concepción rechazan la teoría del Propio Interés. Creen que cuando la moralidad entra en conflicto con el propio interés es irracional seguir el propio interés. Pero otros adoptan la opinión de Sidgwick. Son los que creen que, cuando estos entran en conflicto, sería racional seguir cualquiera de ellos.

Supongamos a renglón seguido que aceptamos alguna forma de Consecuencialismo. Según esta teoría moral, las razones para actuar son neutrales respecto del

Si es así, ¿es este resultado más perjudicial para un bando que para el otro? Muchos suponen que sería más perjudicial para la moralidad. De esta forma, Hume pregunta «¿Qué teoría de la moral puede en absoluto servir a un propósito útil, como no pueda demostrar... que todos los deberes que recomienda constituyen también el verdadero interés de cada individuo?» [3]. Aquí Hume da por sentado que, si se nos forzara a elegir, elegiríamos el propio interés. Como escribe, «¿Qué esperanzas podemos en absoluto tener de ocupar a la humanidad en una práctica que reconocemos llena de rigor y austeridad?». Esta sería nuestra tarea —da él por sentado— si la moralidad y el propio interés pudiesen entrar en conflicto. La solución oficial de Hume es que los dos coinciden siempre [4]. Al poner esto de manifiesto, mostramos a la morali-

agente. No puede afirmarse que el que una persona deba obrar de algún modo depende de cuánto tendría que sacrificarse *ella en particular*. Los intereses del agente no tienen peso especial. No tienen mayor peso que los intereses de cualquier otro. Los consecuencialistas pueden afirmar que, según su teoría, a las demandas del propio interés se les ha dado la importancia debida. Pero esto no puede afirmarse en el sentido en que puede afirmarse para la Moralidad del Sentido Común. Al decidir lo que debe hacer el agente, la Moralidad del Sentido Común le da un peso especial al propio interés del agente. Por tanto, puede afirmarse que le da la debida importancia a la fuerza especial de las razones del propio interés. Los Consecuencialistas no pueden hacer esta afirmación.

Aunque no la pueden hacer, sí que podrían empero rechazar la concepción de que hay conmensurabilidad con pesos diversos. Pueden subrayar que, puesto que C es neutral respecto del agente, aporta razones de una especie muy diferente de las del propio interés. Podrían afirmar entonces que, puesto que estas razones son tan diferentes, no puede ser comparado el peso de unas y otras porque no hay una balanza neutral.

[3] Hume (3), Conclusión, Sección II.

[4] Esta fue casi siempre, antes de Sidgwick, la solución apuntada. Fue la novedad de Hume proclamarla sin apelar a la vida de ultratumba. Llamo «oficial» a la solución de Hume porque sus comentarios pueden ser irónicos. Escribe:

> «Tratando el vicio con la mayor de las franquezas, y haciendo todas las concesiones posibles, tenemos que reconocer que no hay en ningún caso el menor pretexto para darle la preferencia sobre la virtud, en atención al propio interés. Salvo tal vez en el caso de la justicia, donde un hombre, viendo las cosas a una cierta luz, con frecuencia puede parecer un perdedor a causa de su integridad. Y aunque se conceda que sin

dad «en todos su genuinos y más seductores encantos... se cae el deprimente vestido».

Lo que Hume supone es falso. No es verdad que, cuando la moralidad entra en conflicto con el propio interés, todos nosotros elegiríamos siempre el último. Se han dado innumerables casos en que alguien hace lo que cree moralmente correcto, aunque, como no es religioso, piense que está haciendo lo que va a ser peor para él. Si pensamos que, en caso de conflicto, sería racional actuar de cual-

> el respeto a la propiedad ninguna sociedad podría en absoluto subsistir, sin embargo, según el modo imperfecto en que se conducen los asuntos humanos, un bellaco prudente puede pensar en ciertas ocasiones que un acto de iniquidad o infidelidad traerá consigo un considerable aumento para su fortuna... Que la honestidad es la mejor política puede ser una buena regla general, pero está sujeta a muchas excepciones. Y tal vez pueda pensarse que el que se comporta con más sabiduría es aquel que observa la regla general y se aprovecha de todas las excepciones.
>
> Tengo que confesar que, si alguien piensa que este razonamiento necesita urgentemente una respuesta, sería un poco difícil encontrar una que le vaya a parecer a él suficiente.»

¿Están justificadas las palabras *a él*? Podemos dudarlo si recordamos las observaciones anteriores de Hume: «A un hombre no le sucede sino una desgracia si formula una teoría que, si bien verdadera, tiene que confesar que conduce a una práctica peligrosa y perniciosa... ¿Por qué sacar la pestilencia de la fosa en que está enterrada? Las verdades que son *perniciosas* para la sociedad, si es que hay alguna, cederán el paso a errores convenientes y ventajosos». Estas observaciones nos advierten de que lo que Hume afirma después tal vez no sea verdadero. Como escribió Sidgwick, refiriéndose a su juventud: «si me asaltaba una duda respecto de la coincidencia de la felicidad privada con la general, me inclinaba a sostener que debía ser lanzada al viento por una resolución generosa» [Sidgwick (1), p. XV]. Pero Sidgwick no fue capaz después de engañarse a sí mismo ni de engañar al público. Parece poco probable que Hume se engañara a sí mismo.

Tras confesar que no disponía de un razonamiento con el que poder responder al bellaco prudente, Hume escribe: «Si su corazón no se rebela contra esas máximas perniciosas, si no siente rechazo ante los pensamientos de villanía o bajeza...». Tal vez la moralidad no pueda derrotar por sí misma a la teoría del Propio Interés. Tal vez tenga que apelar al corazón y aliarse con la teoría del fin Presente. Si estos aliados derrotan a la teoría del Propio Interés, su alianza terminará. La teoría del fin Presente a menudo entra en conflicto con la moralidad. Pero, si esta es adversaria de la moralidad, la amenaza que representa para ella tal vez sea menor. Espero argumentar esto en otro sitio.

quiera de los dos modos, esta creencia no es, sin duda, más perjudicial para la moralidad.

Sidgwick sabía esto. Después de formular esta creencia, escribió: «la razón práctica, estando dividida contra sí misma, dejaría de ser un motivo en cada uno de los lados: el conflicto tendría que decidirse por la preponderancia comparativa de uno u otro de los dos grupos de impulsos no racionales» [5].

Sidgwick no discutió la teoría del fin Presente. Si esta teoría entra en el campo, y el resultado es un empate a tres bandas, hay una respuesta natural a la pregunta de Sidgwick. Si ninguna de las tres rivales aporta razones más poderosas que las otras dos, dos de las tres podrían, cuando estuvieran de acuerdo, proporcionar razones más poderosas que la tercera. Si es así, podemos hacer la pregunta de qué tenemos más razones de hacer. Supongamos que estoy pensando con claridad, y que sé que será peor para mí si cumplo con mi deber. Si lo que más deseo es cumplir con mi deber, esto es lo que tengo más razones de hacer. Tenemos que admitir que no tendría más razones para tener cualquiera de estos deseos más bien que el otro. Esto se sigue del supuesto de que, cuando la moralidad entra en conflicto con el propio interés, ninguno aporta una razón más poderosa para actuar. Pero aunque mi preferencia por uno de los lados no viniese racionalmente requerida, no sería irracional. Sería, como escribe Sidgwick, «no racional». Y el hecho de que mi preferencia sea no-racional no la hace arbitraria. No sería como escoger al azar una de dos cosas exactamente iguales. Sidgwick pensaba que cuando la moralidad entra en conflicto con el propio interés no podemos tener más razones para seguir ninguno de los dos. Yo he defendido que sí podemos.

Volvamos ahora a la conclusión similar de que, cuando P entra en conflicto con PI, ninguna derrota a la otra. Según esta conclusión, cuando estas teorías entran en conflicto, sería racional seguir cualquiera de las dos.

Afirmo aquí lo que he criticado a Hume por sus afirmaciones sobre el conflicto entre PI y la moralidad. Si el resultado es de

[5] Sidgwick (1), p. 508.

tablas, es más perjudicial para una de estas teorías. Es más perjudicial para PI. Porque, de estas dos teorías, es PI la que necesita ser más crítica. Naturalmente tendemos a hacer lo que sabemos logrará mejor nuestros fines presentes. No estamos seguros de hacer esto. Hay muchos modos en que podemos fallar a la hora de hacer lo que, según la teoría del fin Presente, tenemos más razones de hacer. Pero, si no pensamos que estamos actuando irracionalmente, es esta teoría la que, de las dos, seguiremos con mayor probabilidad.

Puedo actuar de cierta manera porque piense que sería irracional no hacerlo así. En palabras de Sidgwick, la «razón práctica» puede ella misma ser «un motivo». No tenemos que decidir si esto es verdad sólo porque tengo el deseo de no actuar de manera irracional. Desde luego que puedo tener este deseo, y que esto permitiría que mis creencias sobre la racionalidad afectaran a mis actos. Supongamos que puedo o bien (1) hacer lo que mejor logrará lo que, en ese momento, conociendo los hechos y pensando con claridad, quiero o valoro más, o bien (2) hacer lo que vaya a ir a favor de mi propio interés a largo plazo. Y supongamos que creo que, cuando PI y P entran en conflicto, no gana ninguna. Pienso que, en tal caso, sería racional hacer o (1) o (2). Si creo que ningún acto sería irracional, está claro lo que con más probabilidad voy a hacer. Con más probabilidad haré lo que mejor lograse lo que, en el momento de actuar, quiero o valoro más. Entonces no estaré siguiendo a PI sino a P.

Supongamos en vez de eso que pienso que, cuando las dos teorías entran en conflicto, PI derrota a P. En lugar de (PI13) yo acepto la afirmación más contundente (PI12). Podría en ese caso hacer (2) antes que (1). Tengo el deseo de no actuar de forma irracional, y ahora creo que (1) sería irracional, desde el momento en que sería peor para mí. Mi creencia sobre la racionalidad puede así llevarme a seguir PI.

Puede objetarse que, si ahora hago (2) porque considero que sólo (2) es racional, todavía tengo que estar obrando según mi deseo más poderoso, o estar haciendo aquello que realiza mejor mis deseos presentes. Esta afirmación es polémica. Si es verdadera, coincidirían aquí las dos teorías. Pero esto sería verdad sólo porque creo

en la teoría del Propio Interés antes que en la del fin Presente. Las dos teorías entrarían en conflicto si en vez de esto yo creyera en la teoría del fin Presente, o creyera que ninguna de las dos teorías derrota a la otra.

Cuando las dos teorías entran en conflicto, estaremos naturalmente inclinados a hacer lo que vaya a conseguir mejor nuestros fines presentes. Por eso, de las dos, la teoría del Propio Interés es la que necesita hacer la afirmación más contundente. Necesita la afirmación de que sería irracional para mí hacer lo que mejor lograría mis fines presentes, cuando sé que mi acto sería peor para mí. Si pienso que esto sería irracional, puedo entonces se conducido a hacer lo que sería mejor para mí. Pero si considero que ningún acto sería irracional, estaré naturalmente inclinado a hacer lo que mejor vaya a lograr mis fines presentes. Y esto se aplica a todos nosotros. Cuando nuestro deseo de actuar de forma racional, o de evitar la irracionalidad, no nos dice ninguna de las dos cosas, trataremos con más probabilidad de hacer lo que vaya a lograr mejor lo que, en ese momento, queremos o valoramos más. Por esta razón, en su conflicto con la teoría del Propio Interés, lo que teóricamente es un empate, para la teoría del fin Presente es en la práctica una victoria.

C. La racionalidad y las diferentes teorías del Propio Interés

Afirmo (1) que, comparado con la predisposición en nuestro propio favor, hay otros varios deseos o patrones de interés que no son menos racionales. Y concluyo (2) que, si alguien tiene uno de estos otros deseos, no sería menos racional para él actuar a partir de él, aunque sepa que esto va a ir en contra de su propio interés.

Si aceptamos (1) tenemos que aceptar al final (2). Pero ahora describiré cómo, cambiando nuestra concepción del propio interés, podemos posponer la aceptación de (2). Podemos posponer nuestro paso de PI a P.

Describiré una línea de pensamiento que parte de la concepción de Bentham. Si aceptamos alguna otra concepción, entraremos en

esta línea de pensamiento en una etapa posterior. La concepción de Bentham combinaba PI con la Teoría Hedonista del propio interés. Según ella, lo que cada persona tiene más razones de hacer es lo que vaya a hacerle tan feliz como sea posible.

Esta concepción con frecuencia descansaba sobre la versión hedonista del Egoísmo Psicológico: la afirmación de que lo que cada persona más desea, o más desearía en un momento de serenidad, es ser tan feliz como sea posible. Esta afirmación es falsa. Con frecuencia tenemos un deseo distinto, que seguiría siendo nuestro deseo más intenso incluso después de una deliberación serena.

La realización o la búsqueda de satisfacción de tales deseos distintos puede ser nuestra principal fuente de felicidad. Pero como estos son deseos de algo distinto de nuestra propia felicidad, obrar a partir de ellos a veces nos hará menos felices. Esto es muy probable que sea verdad cuando nunca vamos a saber si estos deseos se realizan; pero a menudo es cierto incluso cuando lo vamos a saber.

Según la manera de ver las cosas de Bentham, es irracional en tales casos obrar a partir de estos deseos. Aceptaremos esta afirmación sólo si pensamos que estos deseos son irracionales. Hay personas que lo piensan. Pero, en el caso de muchos deseos como estos, nosotros tal vez estemos en desacuerdo.

Con toda probabilidad estaremos en desacuerdo en el caso del deseo de que otras personas sean felices. Si este fuera el único deseo distinto que creyéramos que no es irracional, estaríamos en desacuerdo con Bentham no sobre la cuestión del propio interés sino sobre la de la racionalidad. Como el objeto de este deseo es todavía la felicidad de alguien, seguiríamos aceptando la Teoría Hedonista del propio interés. Pero le haríamos una matización a PI. Afirmaríamos que no es irracional sacrificar nuestra propia felicidad cuando con ello podemos proporcionar una felicidad mayor a los demás. Esta era la idea de Sidgwick.

Supongamos, a renglón seguido, que pensamos que una gama mucho más amplia de deseos no es irracional. Son los deseos cuyo objeto no es la felicidad de nadie. Una gran clase podría ser la de los deseos que dependen de ciertos juicios de valor, o ideales. Pero la Sección 60 demuestra que, en su tratamiento de estos deseos, PI

tiene que rechazarse. Otra gran clase la denomino *deseos de logro*. Se trata de deseos de tener éxito al hacer lo que tratamos de hacer, ya sea en nuestro trabajo o en nuestro ocio más activo. Así, un artista, un jardinero, un carpintero, un creador de cualquier especie, pueden desear ardientemente hacer su creación lo mejor posible. Su más ardiente deseo puede ser producir una obra maestra, con pintura, flores, madera o palabras. Y un científico o un filósofo pueden desear ardientemente hacer un descubrimiento o un avance intelectual fundamentales.

Deseos como estos pueden tener importancia en sí mismos, y no quedarse en medios para el logro de otros deseos. Puede ser cierto que, si los realizamos, con frecuencia promoveremos nuestra propia felicidad o la de otros. Pero este no es el objeto de estos deseos. Y también hay muchos casos, como la búsqueda de ciertas clases de conocimiento, en que la realización de estos deseos no va a hacer nada para llevar felicidad a los demás. (Es muy normal que también podamos querer no sólo realizar estos deseos sino hacer que nuestros logros se reconozcan. Tal vez deseemos la fama. Pero este es un deseo aparte, que puede ser más débil. Tal vez deseemos reconocimiento sólo porque, sin él, no podríamos tener la seguridad de que realmente habíamos producido una obra maestra o habíamos hecho un descubrimiento fundamental.)

Tratando de realizar estos deseos de logro, en ocasiones nos haremos a nosotros mismos menos felices. Esto puede ocurrir aunque sepamos que estos deseos están realizándose. Por ejemplo, George Eliot sabía que era una novelista de éxito, pero siempre estaba descontenta con lo que había conseguido hasta ese momento [6]. La lucha por el logro puede no ser el tipo de lucha que es su propia recompensa, sino un suplicio.

Supongamos que decidimos que algunos de estos deseos de logro no son irracionales. Concluimos que no es irracional actuar a partir de ellos, aun cuando sepamos que esto nos va a hacer menos felices y no va a proporcionar mayor felicidad a los demás. (Lo que no sería el caso de un novelista de éxito, pero puede suceder con un científico o un filósofo de éxito.) Como hemos decidido que estos actos no son irracionales, tenemos ahora razones para rechazar no sólo la concepción de Bentham, sino también la de Sidgwick. Tenemos razones para rechazar cualquier concepción que se identifique con, o que incluya, la versión hedonista de PI [7].

[6] Véase Haight, de principio a fin.

[7] Puede objetarse: «Tú crees (1) que los deseos de logro no son irracionales, y (2) que, por tanto, no es irracional obrar sobre la base de los mismos, aun sabiendo que supondrá un coste para nuestra propia felicidad. Y afirmas que estas creencias son razones para rechazar el punto de vista de Sidgwick. Pero no es así. Seremos más felices si tenemos poderosos deseos de cosas distintas de nuestra propia felicidad o de la de otras personas. Seremos más felices, globalmente, si tenemos muchos deseos de logro. Esto es cierto aunque estos deseos a veces nos vayan a llevar a ser menos felices. Dados estos hechos, el punto de vista de Sidgwick implica afirmaciones que se corresponden con tus creencias (1) y (2). El punto de vista de Sidgwick implica (3) que es racional para nosotros determinarnos a nosotros mismos a tener, o a mantener, los deseos de logro. E implica (4) que, cuando obramos sobre la base de los mismos, siendo conscientes de que ello supone un coste para nuestra propia felicidad, estos actos no demuestran que seamos irracionales, porque estamos obrando sobre la base de deseos que sería irracional para nosotros determinarnos a perder».

Esta objeción, apuntada por B. Hooker, apela a afirmaciones que hice en el Capítulo 1. Reduce nuestras razones para rechazar el punto de vista de Sidgwick, pero no las elimina. Lo cual quedará claro si recordamos dos de mis otras conclusiones. Sostuve (5) que no se demuestra que nuestros deseos sean racionales por el hecho de que era racional para nosotros determinarnos a nosotros mismos a tenerlos. Y sostuve (6) que no se demuestra que nuestros actos sean racionales por el hecho de que, como obramos sobre la base de tales deseos, no somos irracionales. (5) y (6) quedaron ilustrados en el caso imaginario en que, al ignorar yo tu amenaza, nos hiciste volar a los dos en pedazos. En este caso era racional para mí determinarme a mí mismo a tener el deseo irracional de ignorar tu amenaza. Y cuando yo obré sobre la base de este deseo, mi *acto* fue irracional, aunque podría afirmarse que *yo* no lo fui.

Cuando recordamos (5) y (6), vemos por qué todavía tenemos razones para rechazar el punto de vista de Sidgwick. Este implica que es para nosotros racional determinarnos a nosotros mismos a tener, o a mantener, los deseos de logro. Pero esto no implica nuestra creencia de que estos deseos son racionales. El punto de vista de Sidgwick también implica que, cuando obramos sobre la base de estos deseos, conscientes de que ello representa un coste para nuestra propia felicidad, *nosotros* no somos irracionales. Pero esto no implica nuestra creencia de que estos *actos* no son irracionales. Y, según el punto de vista de Sidgwick, tiene que afirmarse

Hay dos posibilidades. Podríamos cambiar nuestra idea de la racionalidad, yendo de PI a P. La alternativa es cambiar nuestra idea del propio interés. Podríamos irnos de la Teoría Hedonista a la Teoría de la Realización de Deseos.

Para la Teoría de la Realización de Deseos, puede no ir en contra de nuestros intereses hacer lo que nos hace menos felices. La realización de un deseo intenso cuenta ahora directamente como a favor de nuestros intereses, nos haga felices o no. Luchando para resolver su problema, un científico o un filósofo, salvo en raras ocasiones, pueden sentirse desgraciados. Y puede ocurrir que, si lucharan menos duramente, se sentirían menos desgraciados. Puede darse aquí un sacrificio real de su felicidad. Pero si van a realizar su más ardiente deseo, ahora podemos decir que *no* están actuando en contra de sus intereses. Decidimos que no es irracional actuar a partir de tales deseos de logro, aunque nos haga menos felices. Ahora que hemos cambiado nuestra idea del propio interés, esta decisión ya no nos da razones para rechazar PI.

Como pone de manifiesto este ejemplo, si vamos de la Teoría Hedonista a la de la Realización de Deseos, tenemos menos necesidad de ir de PI a P. Esto es porque, una vez dado este paso, PI y P coinciden más a menudo. Pero siguen siendo teorías diferentes. La diferencia esencial concierne una vez más al tiempo. Si asumimos la Teoría de la Realización de Deseos, PI afirma que lo que cada persona tiene más razones de hacer es lo que vaya a realizar mejor, o le vaya a poner en disposición de realizar, todos sus deseos a lo largo de su vida entera. Sus deseos futuros cuentan tanto como sus deseos presentes. Pero P apela sólo a los deseos presentes de una persona —a lo que desea o desearía, tras una deliberación ideal, en el momento de actuar.

PI y P con frecuencia coinciden. Hay muchos casos en que lo que mejor realizaría los deseos presentes de alguien no entraría en conflicto con la realización de sus deseos futuros. Y puede ser cierto que los deseos más intensos de alguien sean los mismos a lo largo de su vida. Pero aunque PI y P coincidan con frecuencia, hay también muchos casos en que entran en conflicto.

que estos actos, como el de que yo ignoré tu amenaza, son irracionales. Como nosotros pensamos que estos actos *no* son irracionales, tenemos razones para rechazar el punto de vista de Sidgwick.

Algunos de estos casos involucran a personas que se preocupan menos por su futuro lejano. Pero ahora estoy suponiendo, para los propósitos del argumento, que hemos condenado la predisposición a favor de lo próximo. Estamos considerando la versión de P que afirma que, en su interés por sí mismo, un agente racional debería preocuparse igualmente por la totalidad de su futuro. Aunque haga esta afirmación, P a menudo entra en conflicto con PI. Esto ocurre sobre todo porque, en el caso de muchas personas, algunos de sus deseos más intensos no duran toda la vida. Los deseos más intensos de muchas personas son de larga duración. Pero hay pocas cuyos deseos más intensos sean siempre los mismos. De este modo mis deseos de logro pueden ser los mismos a lo largo de mi vida, pero puedo en momentos diferentes querer a diferentes personas, y dar mi apoyo a diferentes campañas políticas. O bien puedo querer siempre a las mismas personas y dar mi apoyo a las mismas campañas, pero en diferentes momentos tener diferentes deseos de logro. En cualquier caso, P puede entrar en conflicto con PI. Lo que realizaría mejor mis deseos presentes puede que no coincida con lo que realizaría mejor, o me pondría en condiciones de realizar, la totalidad de mis deseos a lo largo de toda mi vida.

En casos así, PI dice que sería para mí irracional hacer lo que mejor realizase mis deseos presentes. Pero, si estos son deseos de ciertas clases de logro, decidimos que no son irracionales y que no es irracional actuar a partir de ellos. Por consiguiente tenemos nuevas razones para rechazar PI.

Como antes, tenemos dos alternativas. Podríamos cambiar nuestra concepción de la racionalidad, yendo de PI a P. O podríamos de nuevo cambiar nuestro modo de pensar sobre el propio interés. Podríamos ir de la Teoría de la Realización de Deseos a la Teoría de la Lista Objetiva.

De acuerdo con esta teoría, hay ciertas cosas que son buenas o malas para nosotros, sean los que sean nuestros deseos. Una de las cosas buenas pueden ser ciertas clases de logro. Según la Teoría de

la Lista Objetiva, puede ser mejor para mí que realice mi deseo de logro, aunque esto determine que mis deseos a lo largo de mi vida, en conjunto, se vayan a realizar peor. Decidimos que actuar así no es irracional. Ahora que hemos cambiado nuestra opinión sobre el propio interés, esta decisión ya no nos da razones para rechazar PI.

Cuando pasamos de la Teoría Hedonista a la de la Realización de Deseos, el conflicto entre PI y P se redujo pero no se eliminó. No podemos hacer una afirmación tan inequívoca en relación con el paso de la Teoría de la Realización de Deseos a la de la Lista Objetiva. Si damos este paso, esto *cambiará* los casos en que PI entra en conflicto con P. Un ejemplo es el caso que se acaba de dar. Pero no hay razones evidentes para afirmar que, si hacemos este movimiento, habrá *menos* casos en que PI y P entren en conflicto.

Con qué frecuencia entren PI y P en conflicto depende de los contenidos de la Lista Objetiva. La cuestión importante aquí es si, según la Teoría de la Lista Objetiva, PI y P coincidirían siempre. Creo que esto no sería así, según cualquier versión plausible de esta teoría. Para la Teoría de la Lista Objetiva, algunos tipos de logro pueden ser una de las cosas que son buenas para nosotros, y hacen que nuestras vidas vayan mejor. Pero habrá otras varias cosas que sean buenas para nosotros, como por ejemplo el amor mutuo de dos adultos, tener hijos y amarlos, el desarrollo de una gama completa de habilidades y el tomar conciencia de todas las especies de belleza. Puede ocurrir que, para realizar mi deseo de logro, tenga que negarme a mí mismo la mayor parte de las demás cosas que son buenas para mí. De este modo puedo estar haciendo lo que, para la Teoría de la Lista Objetiva, va a ser en conjunto peor para mí. Pero como estoy realizando este deseo de logro, decidimos que no obro irracionalmente. Por tanto, tenemos nuevas razones para rechazar PI.

En dos puntos anteriores de esta línea de pensamiento se nos presentaron alternativas. Podíamos cambiar nuestra idea de la racionalidad o nuestra idea del propio interés. En el punto que hemos alcanzado ahora, *ya no tenemos alternativas*. Ahora sólo hay una conclusión que podamos sacar. No podemos hacer un nuevo cambio en nuestra concepción del propio interés. Tenemos que cambiar nuestra idea de la racionalidad. Tenemos que rechazar PI y aceptar P.

Como dije, si pensamos que hay otros deseos que no son menos racionales que la predisposición en nuestro propio favor, tenemos que rechazar finalmente la teoría del Propio Interés.

Cuando la rechazamos, perdemos el motivo que teníamos para cambiar nuestra opinión sobre el propio interés. Por eso deberíamos reconsiderar estos cambios.

Empecé suponiendo que aceptábamos la Teoría Hedonista, creyendo que va en contra de nuestros intereses ser menos felices. Aquí tenemos un ejemplo diferente. Turner quería que sus cuadros se pudieran ver reunidos en una galería aparte. Supongamos que decidimos que no fue irracional para Turner tratar de asegurar que este deseo se realizaría, incluso con un cierto coste para su propia felicidad. No podemos defender esta creencia mientras aceptemos PI, a no ser que nos vayamos de la Teoría Hedonista a la de la Realización de Deseos. Entonces podemos afirmar que la realización del deseo de Turner iría a favor de sus intereses, aunque le hiciera a él menos feliz. Esto nos permite afirmar que, al tratar de realizar este deseo, con un cierto coste para su propia felicidad, Turner no estaba actuando irracionalmente. Pero estamos forzados a concluir que, al no colocar los cuadros de Turner en una galería aparte, más de un siglo después de su muerte, estamos ahora obrando en contra de los intereses de Turner, o haciendo algo que es malo para él. Y podemos encontrar inverosímil esta afirmación. Tenemos entonces creencias contradictorias.

El conflicto desaparece si abandonamos PI. Ahora podemos afirmar *tanto* que Turner no habría sido irracional si hubiera obrado en contra de sus intereses al tratar de conseguir que hubiera un museo Turner, *como* que, al no construir ese museo, no estamos perjudicando ahora a Turner, o haciendo algo que es malo para él.

Un razonamiento similar, pero con mayor fuerza, se aplica a nuestra elección entre las dos versiones de la Teoría de la Realización de Deseos: la versión No Restringida, y la Teoría del Éxito. No podemos afirmar convincentemente que la realización de *todos* mis deseos vaya a favor de mis intereses. Recordemos el caso en que, después de un viaje en tren, simpatizo con una des-

conocida, y deseo ardientemente que tenga éxito. No es convincente afirmar que si la desconocida después tiene éxito sin yo saberlo, esto es bueno para mí. Es más plausible afirmar que lo que es mejor para mí es sólo que mi propia vida marche del modo que quiero o querría. Si apelamos a esta versión restringida de la Teoría de la Realización de Deseos —la Teoría del Éxito— habrá más casos en que, al actuar a partir de mis deseos presentes, estaré haciendo lo que es peor para mí. Estos serán los casos en que mis deseos no tratan sobre mi propia vida. Si pensamos que algunos de estos deseos no son irracionales, creeremos que no es irracional actuar a partir de los mismos. Si aceptamos PI, podemos así ser llevados a aceptar la Teoría No Restringida de la Realización de Deseos acerca del propio interés. Sólo según ella la realización de estos deseos va a favor de mis intereses. Pero si estos deseos no tratan sobre mi propia vida, y yo nunca sé si se han realizado, podemos encontrar difícil de creer que la realización de estos deseos vaya a ser buena para mí. Podemos tener de nuevo creencias contradictorias.

El conflicto desaparece de nuevo si pasamos de PI a P. Ahora podemos afirmar que no es para mí irracional actuar a partir de estos deseos, aun si hacerlo así va a ser peor para mí. Como hemos rechazado PI, hemos perdido el motivo que teníamos para la inverosímil afirmación de que, si estos deseos no son irracionales, su realización será buena para mí. Hemos perdido el motivo que teníamos para aceptar la Teoría No Restringida de la Realización de Deseos sobre el propio interés. Podemos pasar a la Teoría del Éxito, que es más plausible [8].

[8] Surge ahora una pregunta acerca de la importancia moral de esos deseos del agente de los que juzgamos que no son irracionales, aunque su realización *no* vaya a favor de sus intereses ni de los intereses de nadie. ¿Qué peso moral debemos dar nosotros a tales deseos? Si creemos que debemos dar algún peso a la realización de estos deseos, no podemos apelar a nuestro Principio de Beneficencia corriente, tenemos que apelar a algún otro principio. Podemos afirmar en cambio que ciertos tipos de deseo, mientras que aportan buenas razones para actuar a la persona que los tiene, no aportan razones morales a otras personas. Esta es la idea apuntada en Nagel (3), pp. 121-6.

Nagel cree que lo que él es, esencialmente, es su cerebro. Da tres argumentos en apoyo de esta creencia, argumentos que se contienen en un difícil borrador no publicado que él puede revisar. Aunque el estilo sea difícil los argumentos merecen atención. Por eso citaré de este borrador.

Dos de los argumentos de Nagel apelan a una determinada concepción del significado, la referencia y la necesidad. Los objetos a los que nos referimos tienen algunas propiedades *esenciales*: propiedades que estos objetos tienen que tener porque de lo contrario no podrían existir. Algunas propiedades son esenciales a causa de los significados de nuestras palabras. Así, a causa de lo que significa «triángulo», es una propiedad esencial de los triángulos el tener tres lados. Pero según la manera de pensar que ahora discutiré, hay una forma diferente en que los objetos pueden tener propiedades esenciales. Estas propiedades no son esenciales a causa del significado de nuestras palabras, sino que las *descubrimos* cuando descubrimos hechos sobre lo que es aquello a lo que nos estamos refiriendo. Según esta concepción, por ejemplo, hemos descubierto que una propiedad esencial del oro es tener el número atómico 79. Toda sustancia con este número tiene que ser oro, y ninguna sustancia sin este número podría ser oro. Lo cual no era parte del significado de la palabra «oro» [9].

¿A qué nos estamos refiriendo cuando usamos la palabra «persona» y la palabra «yo»? Nagel escribe: «Lo que yo *soy* es lo que *de hecho* hace posible para la persona TN identificarse y reidentificarse a sí misma y a sus estados mentales». Lo que yo soy es lo que explica la continuidad psicológica de mi vida mental. Y Nagel afirma, de una manera parecida, que esto deja abierto cuál sea la explicación. Si el portador de la continuidad es un Ego Cartesiano, eso es lo que yo realmente soy. Sigue diciendo: «Si por otro lado determinados estados y actividades de mi cerebro subyacen a la capacidad mental,

[9] Véase Kripke, de principio a fin. Esta idea ha sido desarrollada también en varios artículos por H. Putnam.

entonces ese cerebro en esos estados... es lo que yo soy, y no es concebible que yo sobreviva a la destrucción de mi cerebro. Sin embargo, puede que yo no *sepa* que no es concebible, puesto que puedo no conocer las condiciones de mi propia identidad». Como dice más adelante, «al tratar de concebir mi supervivencia tras la destrucción de mi cerebro, no tendré éxito al referirme a *mí mismo* en tal situación si yo soy de hecho mi cerebro».

Nagel es reduccionista. Admite que la identidad personal no conlleva el «hecho adicional», que en todo caso concebible o bien se da completamente o bien no se da en absoluto. La identidad personal nada más que conlleva continuidad física y psicológica. Pero, a pesar de ser reduccionista, el modo de pensar de Nagel difiere en dos aspectos de la concepción que defendemos algunos otros y yo. Según esta lo que fundamentalmente importa es la Relación R: continuidad y conexividad psicológicas. Para el modo de pensar de Nagel, lo que importa es la identidad personal. Y como piensa que él es su cerebro, cree que lo que fundamentalmente importa es la existencia continua de este cerebro.

Puede parecer que esto sólo constituye un desacuerdo en el terreno de los casos imaginarios. No hay casos reales en que haya continuidad psicológica sin la existencia continua del mismo cerebro. Si el desacuerdo fuera sólo sobre casos imaginarios, apenas valdría la pena discutirlo. Pero también incluye casos reales, y nuestras propias vidas. Según mi concepción, una de las dos relaciones que importan, la conexividad psicológica, se da a través del tiempo en grados reducidos. Esta es una premisa esencial de mi argumento, en el capítulo 14, contra la teoría del Propio Interés. Un argumento que se vería socavado si la concepción de Nagel fuera verdadera. La existencia continua del mismo cerebro, en nuestras vidas reales, no es una cuestión de grado.

Esta afirmación requiere una matización. Nagel deja abierta una cuestión importante. Reconsideremos el Ejemplo de Williams, aquel en el que el cirujano manipula mi cerebro hasta que llega a eliminar toda la continuidad psicológica. ¿Sería yo la persona resultante? Aunque hayan manipulado mi cerebro, está claro que se trata del mismo cerebro. Si yo soy mi cerebro, todavía existiré. Pero en

uno de los comentarios que cité arriba, Nagel sugiere que lo que yo soy no es simplemente mi cerebro, sino mi cerebro *en determinados estados*. Quizás sean estos los estados que proporcionan continuidad psicológica. Según esta versión de su concepción, yo no existiría al final del Ejemplo de Williams.

Las dos versiones de la concepción de Nagel podrían reformularse como sigue. Según la versión más simple, lo que yo soy es lo que causa normalmente mi continuidad psicológica. Pero yo sería esta cosa aunque no causara continuidad psicológica. Hemos descubierto que lo que yo soy es mi cerebro. Al final del Ejemplo de Williams el cirujano ha eliminado toda la continuidad psicológica. Pero como mi cerebro todavía existirá, yo todavía existiré.

Según la versión menos simple de la concepción de Nagel, lo que yo soy es lo que causa mi continuidad psicológica, en los estados particulares que hacen que sea esta causa. Para esta versión de la concepción, mi identidad no conlleva sólo la existencia continua de mi cerebro. También lleva consigo continuidad psicológica. Esta versión coincide con una concepción discutida arriba: el Criterio Psicológico Estrecho, que apela a la continuidad psicológica con su causa normal.

Sólo la primera versión del punto de vista de Nagel se halla en desacuerdo en el terreno de los casos reales con la concepción que defiendo yo. ¿Deberíamos aceptar esta primera versión? Depende en parte de si Nagel describe correctamente el significado de las palabras «persona» y «yo». Hay otra complicación. Nagel hace dos afirmaciones sobre lo que él y otros entienden por la palabra «yo». Una es que él usa «yo» con la intención de referirse a lo que explique su continuidad psicológica. La otra es que él usa «yo» con la intención de referirse al «sujeto no observado» de sus experiencias.

Empiezo con la segunda afirmación. Es difícil de negar. Yo no soy una serie de pensamientos, actos y experiencias. Soy el que piensa mis pensamientos y el que hace mis actos. Soy el sujeto de todas mis experiencias, o bien la persona que *tiene* estas experiencias.

Nagel sostiene que, cuando yo uso la palabra «yo», con la intención de referirme a mí mismo, el sujeto de mis experiencias, estoy de hecho refiriéndome a mi cerebro. ¿Deberíamos aceptar esta afirmación?

Primero deberíamos darnos cuenta de que una referencia intentada puede fracasar. Llamemos a eso a lo que estamos intentando referirnos *nuestro referente intencionado*. Puede haber algún objeto que encaje con una de nuestras creencias sobre nuestro referente intencionado. Pero puede que esto no sea suficiente para hacerlo el objeto al que nos estamos refiriendo. Podemos tener muchísimas otras creencias sobre nuestro referente intencionado, que serían falsas en caso de aplicarlas a este objeto. Entonces no nos estaríamos refiriendo a este objeto. Y puede que no nos estemos refiriendo a nada [10].

Un ejemplo sería el siguiente. Los antiguos griegos creían que el dios Zeus era la causa del relámpago y del trueno. Zeus no existía, y la palabra griega «Zeus» no refería a nada. Pero no deberíamos sostener que, como los griegos creían que Zeus era la causa del relámpago y el trueno, y esta causa es un estado eléctrico de las nubes, Zeus es ese estado de las nubes, y esto es a lo que la palabra griega «Zeus» refería. Una estado de las nubes se parece demasiado poco a un dios como para ser el referente de la palabra griega «Zeus».

Al usar la palabra «yo» intento referirme a mí mismo, el sujeto de mis experiencias. Nagel cree que, cuando usamos «yo», la mayoría de nosotros tiene creencias falsas acerca de nuestro referente intencionado. Aunque no seamos conscientes de ello, la mayoría de nosotros piensa que nuestra identidad tiene que ser determinada. Pensamos que somos entidades cuya existencia continua tiene que ser todo-o-nada. Esta creencia habría sido verdadera si cada uno de nosotros hubiese sido un Ego Cartesiano. Pero Nagel piensa que no existen tales entidades. No hay entidades con las propiedades especiales que pensamos que tiene el sujeto de nuestras experiencias. En este sentido, el caso es como el de la palabra griega «Zeus». Pero Nagel sostiene que, al usar «yo», no fracasamos en referir. A lo que

[10] Qué es a lo que nos estamos refiriendo a menudo depende de la historia causal de alguna parte de nuestro lenguaje. Pero no es así siempre, y las consideraciones causales de esta especie no parecen ser de relevancia para la concepción particular que Nagel presenta.

refiere de hecho «yo» es a mi cerebro. Nagel admite que nuestros cerebros no tienen las propiedades especiales que pensamos que tiene aquello a lo que refiere «yo». Pero, mientras que el estado de una nube es demasiado diferente de aquello a lo que los griegos pensaban que refería «Zeus», Nagel afirma que nuestros cerebros no son demasiado diferentes de aquello a lo que creemos que «yo» refiere. Como escribe, este es «uno de esos casos en que una de nuestras creencias más importantes acerca del referente de uno de nuestros conceptos puede ser falsa, sin que de ello se siga que no existe tal cosa».

¿Debemos aceptar esta concepción? Nagel piensa que no somos entidades que existan separadamente, distintas de nuestros cerebros y cuerpos y de nuestras experiencias. Y parece creer que, si la palabra «yo» no refiere a mi cerebro, no hay nada más a que pueda referir. Mi cerebro tiene que ser el sujeto de mis experiencias, puesto que, en la ausencia de Egos Cartesianos, no hay nada más que pudiera ser el sujeto de mis experiencias. Así, tras negar que seamos entidades que existan separadamente, pregunta (1) «¿por qué no recorrer con Parfit todo el camino y abandonar la identificación del yo con el *sujeto* de lo mental...?». Y responde (2) «que el sujeto *real* es lo que importa», aunque no sea el tipo de entidad en que estamos inclinados a creer. (1) da por hecho que, para la Concepción Reduccionista que defiendo, dejamos de creer que *hay* sujetos de experiencias. (2) da por hecho que el sujeto de experiencias es el cerebro.

Por mi parte niego estos dos supuestos. Para la Concepción Reduccionista que defiendo, las personas no son entidades que existan separadamente. La existencia de una persona nada más que implica la existencia de su cerebro y su cuerpo, y el hacer sus actos, y la ocurrencia de sus estados y eventos mentales. Pero aunque no sean entidades que existen separadamente, las personas existen. Y una persona es una entidad que es distinta de su cerebro o cuerpo, y de sus diversas experiencias. Una persona es una entidad que *tiene* un cerebro y un cuerpo, y que *tiene* diferentes experiencias. Mi uso de la palabra «yo» refiere a mí mismo, una persona particular, o sujeto de experiencias. Y yo no soy mi cerebro.

Puede servir de ayuda volver a la analogía de Hume. Podemos ser reduccionistas en lo que se refiere a las naciones, pero creer todavía que las naciones existen, y que nos podemos referir a ellas. Una nación no es una entidad que exista separadamente, algo distinto de sus ciudadanos y la tierra que habitan. La existencia de una nación nada más que consiste en la existencia de sus ciudadanos, actuando juntos de diversas maneras en su territorio. Aunque esto sea todo lo que hay en la existencia de una nación, nosotros podemos hacer referencia a naciones, y afirmar que existen. Así por ejemplo, podemos decir de verdad que Francia existe, y que Francia declaró la guerra a Alemania en 1939. En contraste, no hay ninguna nación llamada Ruritania. Podemos hacer las mismas afirmaciones sobre las personas. Algunas personas existen, y nos podemos referir a ellas, mientras que otras no existen nunca y no nos podemos referir a ellas. Thomas Nagel y yo somos dos de las personas que de hecho existen, y a las que podemos hacer referencia. Pero no nos podemos referir a mi no existente antepasado romano, *Teodoricus Perfectus*.

Mi afirmación siguiente tiene una importancia especial en esta discusión. Cuando usamos la palabra «Francia» para referirnos a una nación, no nos estamos refiriendo a nada distinto de una nación. No nos estamos refiriendo al gobierno de esta nación, ni a sus ciudadanos ni a su territorio. Lo cual puede demostrarse como sigue. Si «Francia» refiriera al gobierno francés, Francia dejaría de existir si el gobierno dimitiera y hubiera un período de anarquía. Pero esto es falso. Las naciones siguen existiendo durante períodos en los que no tienen gobierno. De manera similar, si «Francia» refiriera en 1939 a los que entonces eran ciudadanos franceses, Francia dejaría de existir cuando estos ciudadanos dejen de existir. También esto es falso. Y si «Francia» refiriera a estos ciudadanos, tienen que haber sido ellos los que declararon la guerra a Alemania. También esto es falso. *Hay* un uso de la palabra «Francia» que no refiere a la nación sino al país, o al territorio de esta nación. Cuando decimos que Francia es hermosa, nos estamos refiriendo a su tierra y sus edificios. Pero, según el otro uso, «Francia» refiere a la nación, no a su territorio. Si «Francia» refiriera al territorio francés, Francia no

podría dejar de existir si su territorio no dejara de existir. También esto es falso. Lo que una vez fue el territorio de la nación Prusia existe aún, pero Prusia ha dejado de existir.

Como revela el caso de las naciones, podemos hacer referencia a algo aunque no sea una entidad que existe separadamente. Y al hacer referencia a cosa semejante no nos estamos refiriendo a las otras diversas entidades que están involucradas en su existencia. Si somos reduccionistas acerca de las personas, podemos hacer afirmaciones similares sobre nuestro uso de la palabra «yo». Esta puede referir a una persona, o sujeto de experiencias, aunque una persona no sea una entidad que exista separadamente. Y cuando decidimos que una persona no es una entidad que exista separadamente, no estamos obligados a concluir que una persona tiene que ser o su cerebro o su cuerpo completo. Aunque las naciones no sean entidades que existen separadamente, no estamos obligados a concluir que una nación tiene que ser ni su gobierno ni sus ciudadanos ni su territorio ni los tres juntos. Una nación no es ninguno de estos tres. Y podemos referirnos a naciones. De forma parecida, no estamos obligados a concluir que una persona sea su cerebro ni su cuerpo completo. Y podemos referirnos a personas [11].

Coincido con Nagel en que la mayoría de nosotros tiene falsas creencias sobre el referente intencionado de la palabra «yo». La mayoría de nosotros piensa que somos entidades cuya existencia continua tiene que ser todo-o-nada. Puede objetarse que, como no hay entidades semejantes, debemos concluir que, tal y como la mayoría la usamos, la palabra «yo» fracasa a la hora de hacer referencia, exactamente igual que fracasa «Zeus». Como Nagel, estoy en disposición de rechazar esta afirmación. «Zeus» no refiere porque el estado de una nube es demasiado poco parecido a Dios. Para la Concepción Reduccionista, las personas son distintas de las personas para la Concepción No-Reduccionista. Pero son mucho más parecidas que los dioses y los estados de las nubes. Por eso puedo afirmar que las personas existen. Y como las personas existen, aun-

[11] Véanse las observaciones de Kripke citadas en la nota final 11 a la Tercera Parte.

que en una forma diferente de aquella en la que estamos inclinados a creer, nuestros intentos de referirnos a las personas puede decirse que tienen éxito. Como Nagel, podemos decir que este es uno de los casos en que tenemos falsas creencias acerca de nuestro referente intencionado, sin que de ello se siga que no nos estamos refiriendo a esta cosa.

Como yo soy una persona, que existe, parezco ser el mejor candidato para eso a lo que mi uso de «yo» se refiere. Nagel podría replicar como sigue. Es cierto de la mayoría de nosotros que creemos ser entidades que existen separadamente. Según la Concepción Reduccionista, esta creencia es falsa. Pero esta creencia sería verdadera cuando la aplicásemos al cerebro de una persona. Puede decirse que esto convierte al cerebro de una persona en un *candidato mejor* para aquello a lo que se refiere su uso de «yo».

Hay algo en esta afirmación, pero creo que no es suficiente. Si uso la palabra «X» tratando de referirme al objeto llamado X, y X existe, la asunción natural es que me refiero a X. Pero esta asunción podría no estar justificada si fuera verdadero tanto que X carece de la *mayoría* de las propiedades que creo que tiene, como que algún otro objeto Y tiene la *mayoría* de estas propiedades. Esto podría justificar la afirmación de que, aunque estoy tratando de referirme a X, de hecho me estoy refiriendo a Y. Cuando alguien usa la palabra «yo», su referente intencionado —aquello a lo que está tratando de referirse— es él mismo. Esta persona puede pensar que es una entidad que existe separadamente, distinta de su cerebro y de su cuerpo. Entonces sería verdadero, como afirma Nagel, que el referente intencionado de esta persona carece de una de las propiedades que ella cree que tiene, mientras que otra entidad, su cerebro, tiene esta propiedad —ser una entidad que existe separadamente (aunque no, desde luego, distinta de él mismo)—. Pero el cerebro de una persona no tiene la mayor parte de las propiedades que la mayoría de nosotros pensamos tener. No tiene, por ejemplo, una existencia continua que tenga que ser todo-o-nada, ni tampoco es indivisible. Cuando la mayoría de nosotros usa la palabra «yo», nuestros cerebros no son muy parecidos a lo que pensamos que son nuestros referentes intencionados. Esto va contra la afirmación de que, cuando usamos «yo», nos estamos refiriendo de hecho a nuestros cerebros. Lo que va a favor de esta afirmación es que nuestros cerebros tienen una de las propiedades que erróneamente creemos que tenemos nosotros, la de ser entidades que existen separadamente. Pero pienso que esto no basta para justificar la afirmación de Nagel. No basta para hacer falsa la respuesta natural de que nos estamos refiriendo a nuestro referente intencionado. Cuando usamos «yo» no estamos tratando de referirnos a nuestros cerebros sino a nosotros mismos. Nuestros cerebros tienen una propiedad que erróneamente creemos que tenemos nosotros mismos, pero esto no basta para demostrar que, cuando tratamos de referirnos a *nosotros mismos*, fracasamos. Podemos conservar nuestra creencia natural de que *podemos* referirnos a nosotros mismos.

El segundo argumento de Nagel apela a su otra afirmación sobre el significado de la palabra «yo». Esta es la afirmación de que, al usar «yo», tenemos la intención de referirnos a lo que explique nuestra continuidad psicológica. Creo que podemos rechazarla. Hay aquí un contraste entre los argumentos de Nagel. Cada uno implica una afirmación sobre el significado, y otra sobre los hechos. Yo acepto la afirmación de que uso «yo» con la intención de referirme al sujeto de mis experiencias. Pero he negado la afirmación de que el sujeto de mis experiencias es mi cerebro. Al considerar *este* argumento a favor de la Concepción de Nagel, acepto su afirmación sobre el significado pero niego la que hace sobre los hechos. Al considerar su otro argumento acepto su afirmación sobre los hechos. Lo que explica mi continuidad psicológica es la existencia continua de mi cerebro. Pero niego su afirmación sobre el significado, la de que uso «yo» con la intención de referirme a lo que explica mi continuidad psicológica.

El tercer argumento de Nagel apela a un caso imaginario en el que lo que le parece que importa es la supervivencia de su cerebro. Describe un caso como el del Teletransporte. En él, muchos aceptarían la afirmación de Nagel. Pensarían que lo que importa es la supervivencia de sus cerebros.

Ahora describiré dos casos en los que esto es más difícil de creer. Recordemos antes que un objeto puede seguir existiendo aunque todos sus componentes sean sustituidos. El ejemplo típico es el del barco, del que se ha sustituido después de cada viaje un trozo de madera. Podríamos pensar lo mismo de un cerebro. Hemos aprendido que las células del resto del cuerpo se sustituyen todas gradualmente. Aunque no es verdadero de nuestro cerebro, podría haberlo sido. Nuestros cerebros podrían haber seguido existiendo aunque, como el resto de nuestro cuerpo, tuviesen reemplazados todos sus componentes de manera natural y gradual. Y podemos pensar que nuestros cerebros seguirían existiendo si nosotros mismos causásemos tal recambio gradual. Asumiré aquí que esto es verdadero. Como pensamos que el resto del cuerpo de la persona sigue existiendo aunque sus componentes sean sustituidos gradualmente, ¿por qué deberíamos adoptar una opinión diferente acerca de nuestros cerebros?

A continuación supongamos que necesito cirugía. Todas mis células cerebrales tienen un defecto que, con el tiempo, sería mortal. Pero un cirujano puede reemplazarlas a todas. Puede insertar células nuevas que son réplicas exactas de las células existentes, pero que no tienen ese defecto. Podemos distinguir dos casos.

En el *Caso Uno*, el cirujano realiza cien operaciones. En cada una de ellas, elimina una centésima parte de mi cerebro e inserta una réplica de esta parte. En el *Caso Dos*, el cirujano sigue un procedimiento diferente. Primero elimina todas las partes de mi cerebro, y luego inserta todas su réplicas.

Hay una diferencia real entre estos casos. En el Caso Uno, cada nueva parte de mi cerebro se une durante un tiempo al resto de él. Esto permite a cada nueva parte convertirse en parte de mi cerebro. Cuando la primera parte nueva se inserta, y se une al resto de mi cerebro, gana el título a ser tan parte de mi cerebro como las partes antiguas. Cuando la segunda parte nueva se inserta, también se convierte en una parte de mi cerebro. Esto es verdadero de cada nueva parte, porque hay un tiempo en que esta parte se une a lo que entonces cuenta como el resto de mi cerebro.

En el Caso Dos, las cosas son diferentes. No hay espacios de tiempo en los que cada nueva parte se una al resto de mi cerebro. Por eso las partes nuevas no cuentan como partes de mi cerebro. Mi cerebro deja de existir.

Algo parecido podría ser verdadero en relación con la existencia de un club. Consideremos un club que está limitado a cincuenta miembros. Todos los miembros que lo forman quieren renunciar. Y otras cincuenta personas quieren entrar en el club. Hay una regla que dice que no se puede admitir a un nuevo miembro a no ser en presencia de cuarenta nueve miembros efectivos. A causa de esta regla, el club sigue existiendo sólo si lo que ocurre es como el Caso Uno. Lo que ocurre tiene que ser esto. Un miembro dimite y un nuevo miembro es admitido. Un tercer miembro dimite y un nuevo miembro es admitido. Al término de esta serie, el club todavía existiría, con miembros completamente nuevos. Supongamos en cambio que lo que ocurre es como el Caso Dos. Todos los miembros antiguos renuncian. A causa de la regla, los nuevos miembros no pueden ser admitidos. El club deja de existir.

Volvamos a los Casos Uno y Dos. Estoy asumiendo que un cerebro, por un proceso de sustitución gradual, podría llegar a estar compuesto de nuevos componentes. Según esta suposición, está claro que, en el Caso Uno, mi cerebro sigue existiendo, y que, en el Caso Dos, no. Nagel sugiere que la identidad es lo que importa, y que yo soy mi cerebro. Según este modo de pensar, el Caso Uno me da la vida y el Caso Dos la muerte.

¿Es esto creíble? Aunque haya una diferencia real entre estos dos casos, es menor que la diferencia que Nagel tenía en mente. Él consideraba un caso en que su cerebro sería destruido, y una Réplica creada. Y él comparaba este caso con la supervivencia corriente, en que su cerebro sigue teniendo todas las mismas células existentes.

En mi par de casos, la diferencia es más pequeña. En ellos, más tarde habrá una persona cuyo cerebro será exactamente como mi cerebro actual, quitando los defectos. Como resultado, esta persona será completamente continua conmigo desde el punto de vista psicológico. Y, en *ambos* casos, el cerebro de esta persona estará compuesto de *los mismísimos* componentes nuevos, cada uno de los

cuales es una réplica de una parte de mi cerebro. La diferencia entre los casos es simplemente la manera en que estas partes nuevas son insertadas. Es una diferencia en el orden de eliminaciones e inserciones. En el Caso Uno, el cirujano alterna entre eliminar e insertar. En el Caso Dos, hace toda la eliminación antes de la totalidad de la inserción.

¿Puede ser *esta* la diferencia entre la vida y la muerte? ¿Puede depender *mi* destino de esta diferencia en el orden de las eliminaciones y las inserciones? ¿Puede ser tan importante para mi supervivencia el que las partes nuevas se unan durante un tiempo a las partes antiguas? Esto podría representar toda la diferencia si produjera un hecho adicional. Esto sería así si mi supervivencia fuese como un poder sagrado que un sacerdote pudiera darle a otro sólo a través de un ritual que implicase contacto. Pero no hay tal hecho adicional. Hay simplemente el hecho de que, si las partes nuevas son unidas durante un tiempo a las antiguas, describimos el cerebro resultante como el mismo cerebro. Si las partes nuevas no son unidas así, describimos el cerebro resultante como un cerebro diferente.

Nagel no cree que las razones a favor de su modo de pensar sean decisivas. Y admite que «es difícil *internalizar* una concepción de mí mismo como idéntico a mi cerebro». Él adopta esta idea en parte porque, en el par de casos que consideró, su supervivencia le parecía depender de si su cerebro seguía existiendo. En mi pareja de casos, la diferencia en lo que ocurre es mucho menor. Si considerara estos casos, Nagel podría cambiar su forma de pensar. Él propone tanto (1) que la identidad es lo que importa, como (2) que él es su cerebro. Pero admite que (2) es difícil de aceptar. Es difícil de creer que uno mismo es su cerebro. Cuando considero los Casos Uno y Dos, encuentro imposible creer tanto en (1) como en (2). No puedo creer que lo que importaría para mi supervivencia es que, a lo largo de un período de tiempo dado, las réplicas de partes de mi cerebro fueran insertadas de una de estas dos maneras. No puedo creer que si el cirujano alterna la eliminación y la inserción será tan bueno como la supervivencia corriente, mientras que, si efectúa toda la eliminación antes de proceder con toda la inserción, será casi tan malo como la muerte corriente. Si no es lo que realmente

importa esta diferencia entre los dos casos, hay dos alternativas. O la identidad no es lo que importa o yo no soy mi cerebro.

Le presta apoyo a la primera alternativa el caso imaginario en que me divido. En este caso, cada una de las dos personas resultantes tiene la mitad de mi cerebro. Y no hay replicación. Estas mitades estarán compuestas de mis células cerebrales existentes. Como he defendido, es muy difícil de creer que yo debería considerar la división como equivalente a la muerte. Mi relación con cada una de las personas resultantes contiene todo lo que se necesitaría para la supervivencia corriente. Y esto sigue siendo verdadero *aunque lo que yo soy sea mi cerebro*. Cada mitad de mi cerebro seguirá existiendo y manteniendo vida consciente. Cada una de las personas con la mitad de mi cerebro dará la impresión de recordar toda mi vida, y será en todo otro aspecto continua conmigo desde el punto de vista psicológico. Si soy mi cerebro, este no es un caso en que muero porque mi cerebro deja de existir. Mi cerebro sigue existiendo, y, puesto que está dividido, mantiene la vida con más abundancia. No sólo mantiene una, mantiene dos vidas.

Supongamos que Nagel está de acuerdo en que mi relación con cada persona resultante contiene lo que importa. ¿Podría defender entonces su supuesto de que la identidad es lo que importa? Hemos visto que sí podría, siempre y cuando distorsionara hasta el absurdo nuestro concepto de persona. Doy por sentado que Nagel rechazaría estas distorsiones. Él estaría de acuerdo en que, después de que yo me divida, no habrá nadie viviendo que sea yo. Si también admite que mi relación con cada persona resultante es tan buena como la supervivencia, tendrá que renunciar a la asunción de que es la identidad lo que importa. Sin esta asunción, no estoy obligado a concluir que la replicación sería tan mala como la muerte. Puedo estar de acuerdo en que mi Réplica no es yo, pero afirmo que mi relación con ella contiene lo que fundamentalmente importa. Puedo afirmar que lo que importa es la conexividad y/o la continuidad psicológicas, con cualquier causa.

Nagel podría contestar como sigue. Podría admitir que hay un caso especial en que la identidad no es lo que importa. Es el caso en que dos personas futuras tienen cada una la mitad de mi cerebro.

Pero lo que aquí importa es simplemente la existencia continua de mi cerebro dividido. En todos los demás casos, la identidad personal es lo que importa.

Esta contestación puede tener alguna fuerza. Pero si creemos que la identidad es lo que importa, es natural creer también que la identidad es lo que importa *siempre*. Si admitimos una excepción, puede ser difícil de justificar que rechacemos otras. Dada la pequeña diferencia entre mis Casos Uno y Dos, podemos decir que, también aquí, la identidad no es lo que importa. Si esta afirmación está justificada en el caso en que me divido, ¿por qué no puede estar justificada aquí? Es difícil de creer que mi destino dependa de la diferencia entre estos casos. A diferencia del par de casos que Nagel consideró, esta pareja parece indicar que yo no soy mi cerebro.

He tratado de responder a los argumentos de Nagel. Mis respuestas no demuestran que su concepción sea falsa. Pero creo que demuestran que podemos adoptar una concepción diferente, y hacerlo de una manera defendible. La cuestión sigue abierta. Por eso, en la Sección 98, ofrezco una respuesta muy diferente a la concepción de Nagel.

E. El Esquema del Continuador más Directo

Nozick presenta una concepción general acerca de todos los juicios de identidad a través del tiempo. Ser la misma cosa que una cosa pasada es ser *el continuador más directo* de esa cosa. La concepción de Nozick es reduccionista. Él afirma que puede haber diversas clases de continuidad entre una persona pasada y varias personas presentes. La persona presente que juzgamos que *es* esta persona pasada es la persona presente que tiene la mayor continuidad con esta persona pasada. Según este modo de ver, el hecho de la identidad personal a través del tiempo nada más que consiste en el darse de tales continuidades. No implica ningún Hecho Adicional. Y Nozick compara explícitamente la identidad personal con la identidad a través del tiempo de cierto grupo de filósofos, el Círculo de Viena. Es como mi comparación con la historia de un club.

Aunque la concepción de Nozick sea reduccionista en estos aspectos, él rechaza la versión del Reduccionismo que yo defiendo. Como escribe:

> «Una aproximación filosófica a un área enmarañada de complicadas relaciones de grados diversos, más que tratar de forzarlas a entrar dentro de casillas en cierto modo arbitrarias, se queda contenta limitándose a reconocerlas y a delinear las complicadas relaciones subyacentes. En lo que concierne a la identidad personal, se podría decir que los yoes futuros tendrán diversos grados de proximidad con nosotros-ahora, en virtud de diversas relaciones y eventos subyacentes, tales como la continuidad corporal, la similitud psicológica, la fisión o la fusión; y que toda la verdad real que hay que contar es acerca de la existencia y los contornos de estos fenómenos subyacentes. ¿Por qué imponer cualquier categorización —y la del esquema del continuador más directo es una— a esta complejidad?».

Luego rechaza explícitamente mi concepción, para continuar diciendo:

> «En el nivel subyacente mismo también surgirán problemas parecidos. Por ejemplo, ¿en qué sentido es algo el mismo cuerpo cuando todas sus células diferentes de las neuronas, así como las moléculas especiales que componen las neuronas, se sustituyen a través del tiempo? ¿Deberíamos una vez más hablar sólo de las complicadas relaciones que subyacen a este nivel? No podemos eludir el esquema del continuador más directo, o alguna otra categorización, por el procedimiento de limitarnos a toda la complejidad de las relaciones subyacentes... Al final somos empujados a un esquema del continuador más directo o a algo similar a un nivel u otro... Si se hace legítimo, por necesario, usar el esquema en algún nivel, entonces ¿por qué no empezar con él, simplemente?» [12].

[12] Nozick (3), pp. 60-1.

Este desacuerdo es innecesario. Yo no niego que hagamos juicios sobre la identidad a través del tiempo de muchas clases diferentes de cosas. Y acepto el Esquema del Continuador Más Directo de Nozick como una explicación de cómo hacemos muchos de tales juicios. Mis tesis son estas. Como la identidad personal a través del tiempo nada más que consiste en el darse de ciertas otras relaciones, lo que importa no es la identidad sino algunas de estas otras relaciones. Y la lógica de la identidad no siempre coincide con lo que importa. Cuando lo que importa adopta una forma ramificada, o se da en grados intermedios, los juicios de identidad no pueden hacerse verosímilmente para corresponder con lo que importa. En estos casos no deberíamos aplicar el Esquema del Continuador Más Directo en un intento de lograr tal correspondencia. Como escribe Nozick, entonces estaríamos forzando a lo que importa «a entrar dentro de unas casillas en cierto modo arbitrarias». En estos casos simplemente deberíamos describir las maneras en que, y los grados en que, se dan estas otras relaciones. Entonces deberíamos tratar de decidir cuánto, y de qué formas diferentes, importan estas relaciones. La objeción de Nozick a este modo de ver las cosas, citada arriba, parece ser que no podemos evitar hacer algunos juicios sobre la identidad. Pero esta es una objeción sólo si añadimos la afirmación de que, si describimos algunos casos haciendo o negando juicios sobre la identidad, *tenemos* que describir *todos* los casos de esta manera. No puedo ver ninguna razón para aceptar esta afirmación.

En su afirmación sobre lo que importa, Nozick parece de nuevo rechazar el Reduccionismo. Dice que, según su opinión, «me preocuparé por igual por mi continuador más directo, cualquiera que sea su grado de proximidad (con tal de que sea lo suficientemente próximo)». Si considerara mi Espectro Combinado, Nozick probablemente retiraría esta afirmación. En los casos que se hallan en la mitad de este Espectro, habría una persona resultante que tendría alguna proporción de las células de mi cerebro y de mi cuerpo, y que en determinados aspectos sería psicológicamente continua conmigo como soy ahora. Pero esta persona resultante también tendría muchas células distintas nuevas, y él o ella también sería de muchas maneras psicológicamente continua con Greta Garbo. En el extremo lejano de este Espectro, esta persona futura no estaría relacionada conmigo de ninguna manera. Si yo aceptara la concepción de Nozick, me preocuparía por igual por tal persona futura, con tal de que él o ella estuviera relacionada conmigo de una manera lo suficientemente directa. Considero que todos estos casos en la primera parte de este Espectro son igual de buenos que la supervivencia corriente. A medida que nos movamos por el Espectro, la persona futura estará menos y menos directamente relacionada conmigo. Pero yo estaría igual de preocupado por esta persona, con tal de que el grado de proximidad fuese lo *suficientemente* próximo.

Según esta concepción, tengo que decidir exactamente qué grado de proximidad cuenta como suficientemente próximo. Una vez más, tengo que trazar una línea divisoria nítida en este Espectro. Si mi relación con una persona futura se halla justo en el lado cercano de esta línea, es tan buena como la supervivencia corriente. Si mi relación con una persona futura se halla justamente más allá de esta línea, yo debería estar menos preocupado. Pero la persona futura en el segundo de estos casos apenas sería en absoluto diferente de la persona en el primer caso. Las diferencias serían sólo que serían reemplazadas unas pocas células más, y que habría un pequeño cambio psicológico, por ejemplo un nuevo deseo de estar solo. Aunque las únicas diferencias sean estas, yo debería preocuparme menos por lo que le ocurre a esta segunda persona.

Este patrón de intereses me parece irracional. ¿Cómo puede tener semejante importancia el que simplemente unas pocas células más sean reemplazadas, o el que haya un pequeño cambio psicológico más? La concepción de Nozick trata este Espectro como si involucrara, en algún punto, una discontinuidad. Pero esto es falso. Puesto que el Espectro es llano, y da cabida a todos los grados de continuidad, ¿por qué preocuparnos por igual en todos los casos de la primera parte del Espectro y luego pasar de repente a preocuparnos menos? Esto sería racional sólo si la identidad fuese un hecho adicional que se diera completamente en la primera parte de este espectro y luego dejara de darse repentinamente. Pero Nozick no cree que haya ningún hecho adicional semejante.

Nozick apunta otra manera en que se podría defender su patrón de intereses. Piensa que puede ser racional adoptar lo que denomina el *modo platónico* de interesarse por algo. En este modo, «vemos el mundo en su aspecto de realizar lo que se halla más allá de él, vemos y podemos responder a los destellos de algo más perfecto que brilla a su través». Aunque no es verdad que seamos seres cuya existencia continua tenga que ser todo-o-nada, puede ser racional preocuparse por nuestra identidad como si esto fuera verdad. Como admite Nozick, esto implica «una sobreestimación no realista de la realidad, verla con gafas platónicas». La alternativa es hacer «una valoración más realista de las cosas, verlas como son en sí mismas». Nozick objeta a esta alternativa que «le hace a uno prisionero o víctima del mundo real, limitado por sus insuficiencias, por cómo da la casualidad que es...» [13].

¿Es esta una defensa suficiente del patrón de intereses de Nozick? ¿Puede ser racional para sus intereses no estar en correspondencia con la verdad real de su vida, sino con lo que a él le hubiera gustado que fuera la verdad? Supuesta la distinción entre racionalidad teórica y racionalidad práctica, el patrón de intereses de Nozick es defendible. Por eso retiro la objeción que puse arriba. Una vez más, no hay desacuerdo. Si Nozick reacciona a la realidad no como ella es sino como le hubiera gustado que fuese, esto es irracional desde el punto de vista teórico. Pero si esta clase de pensamiento desiderativo resulta más profundamente satisfactorio, puede ser para él racional desde el punto de vista práctico tratar de hacerse a sí mismo, de este modo, irracional desde el punto de vista teórico.

F. La Tasa de Descuento Social

De acuerdo con una Tasa de Descuento Social, la importancia moral presente de los sucesos futuros, especialmente beneficios y pérdidas, disminuye a una tasa de *n* por ciento por año. Dos tasas comúnmente empleadas son las del 5 y las del 10 por ciento. Contra una clase de *TDS* no tengo nada que decir. Es la TDS aplicada a beneficios y pérdidas medidos en términos monetarios, para el supuesto de que habrá inflación. Pero muchos economistas aplican una TDS a beneficios y pérdidas medidos con el tamaño que tendrán cuando ocurran. Mi ejemplo en el texto no es imaginario. Se ha sugerido totalmente en serio que al valorar los riesgos de la eliminación de los desechos nucleares, deberíamos aplicar una TDS a las *muertes* futuras. En un sentido más general, las Tasas de Descuento Social se han aplicado no a las ganancias y a las pérdidas monetarias sino a lo que los economistas llaman la utilidad real que disfrutarán las personas futuras. Esta es la clase de TDS que discuto. ¿Por qué se ha pensado que esta clase de TDS está justificada? Sé de seis argumentos.

[13] Nozick (3), pp. 67-8.

El Argumento de la Democracia: Muchas personas se preocupan menos del futuro más distante. Hay autores que dicen que, si esto pasa con la mayor parte de los ciudadanos adultos de un país democrático, el gobierno de este país deberá emplear una Tasa de Descuento Social. Si a su electorado le importa menos el futuro más distante, un gobierno democrático deberá hacerlo así. En caso contrario sería paternalista o autoritario. Para decirlo con las palabras de un autor, las decisiones del gobierno deberían «reflejar sólo las preferencias de los individuos presentes» [15]. Podemos ignorar este argumento. Tenemos dos preguntas:

(1) Como comunidad, ¿podemos usar una Tasa de Descuento Social? ¿Estamos moralmente justificados en preocuparnos menos por los efectos más remotos de nuestras políticas sociales, a una tasa de *n* por ciento por año?
(2) Si la mayor parte de nuestra comunidad contesta «Sí» a la pregunta (1), ¿deberá nuestro gobierno hacer caso omiso de esta opinión mayoritaria?

El Argumento de la Democracia se aplica sólo a la pregunta (2). Para la pregunta (1), que es lo que nos interesa, el argumento es irrelevante.

[15] Marglin.

La cuestión se puede poner en estos términos. Un demócrata cree en ciertos acuerdos constitucionales. Estos le proporcionan su respuesta a la pregunta (2). ¿Cómo podría darle una respuesta a la pregunta (1) su compromiso con la democracia? Sólo si asume que lo que la mayoría desea, o cree que es correcto, tiene que *ser* correcto. Pero ningún demócrata sensato asume esto. Supongamos que una mayoría quiere emprender una guerra de agresión, sin importarle nada la matanza de extranjeros inocentes. Esto no demostraría que tienen razón en no preocuparse. Del mismo modo, aunque la mayoría de nosotros nos preocupamos menos por los efectos más remotos de nuestras políticas sociales, y pensamos que este interés menor se halla justificado moralmente, esto no puede demostrar que *esté* justificado. Sea lo que sea lo que la mayor parte de nosotros desee o piense, esta cuestión moral sigue abierta.

Puede objetarse: «En algunos casos, no es esta una cuestión moral. Supongamos que en un referéndum votamos a favor de una política social que nos afectará sólo a nosotros. Y supongamos que, porque nos preocupamos menos por lo que nos ocurrirá más adelante, votamos a favor de una política que nos reportará beneficios ahora, al coste de mayores cargas más adelante. Esta política va en contra de nuestros intereses. Pero como sólo nos afectará a nosotros, no podemos estar actuando inmoralmente al votar por ella. Como mucho podemos estar obrando de manera irracional».

Según los supuestos que la mayoría de nosotros acepta, estas afirmaciones aportan alguna defensa a la Tasa de Descuento Social. Pero la defensa rara vez se aplica. La mayoría de las políticas sociales les afectarán a nuestros hijos además de a nosotros mismos. Si votamos por una política que irá en contra de los intereses de nuestros hijos, la mayor parte de nosotros admitiría que esto sería, en algún grado, moralmente incorrecto. Observaciones similares se aplican a los intereses de las personas que todavía no han nacido. Es una cuestión moral la de cuánto peso debemos dar a sus intereses. Otra objeción es que los votos rara vez son unánimes. Si una política va a ir en contra de nuestros intereses, probablemente irá en contra de los intereses de la minoría que votó en su contra. La mayoría, que votó a su favor, estaría obrando

entonces contra los intereses de esta minoría. Esta es otra razón para la crítica moral.

Como ponen en evidencia estos comentarios, habría pocos casos en que la Tasa de Descuento Social no levantara una cuestión moral. Si una de mis afirmaciones anteriores estuviera justificada, no habría tales casos. Yo defendí que, si nos preocupamos menos por lo que nos sucederá más adelante, y, por tanto, obramos en contra de nuestros propios intereses, puede que esto no sea irracional. Pero tales actos se hallan abiertos a la crítica. Afirmé que deberíamos considerarlos moralmente incorrectos. Si esto es así, la Tasa de Descuento Social plantea siempre una cuestión moral.

> *El Argumento de la Probabilidad.* Con frecuencia se afirma que deberíamos descontar los efectos más remotos porque son los que van a ocurrir con menor probabilidad.

Aquí hay dos preguntas:

(1) Cuando una predicción se aplica al futuro más distante, ¿es menos probable que sea correcta?
(2) Si una predicción es correcta, ¿le podemos dar menos peso porque se aplique al futuro más distante?

La respuesta a (1) es con frecuencia Sí. Pero esto no aporta ningún argumento para contestar Sí a (2).

Supongamos que estamos decidiendo si abandonamos o incrementamos nuestro uso de energía nuclear. Consideramos posibles accidentes, con cálculos de muertes previstas a causa de escapes radiactivos. En un accidente pequeño, esas muertes podrían todas confinarse a lo estadístico, en el sentido de que nosotros nunca sabríamos qué muertes particulares había causado este accidente. Cuando consideramos posibles accidentes, tenemos que pensar en el futuro lejano, puesto que hay elementos radiactivos que siguen siendo peligrosos durante miles de años. De acuerdo con una Tasa de Descuento Social del 5%, una muerte estadística el año próximo cuenta por más de mil millones de muertes dentro de 400 años. Comparado con causar una sola muerte, es moralmente menos

importante que la política que hemos elegido cause mil millones de muertes. Esta conclusión es disparatada. Los mil millones de persones morirían en un futuro muy lejano, pero esto no puede justificar la afirmación de que, en comparación con matar a una sola persona, estaríamos obrando menos mal si en vez de eso matásemos a mil millones de personas. El Argumento de la Probabilidad no lleva a esta conclusión. Como mucho podría llevar a una conclusión diferente. Sabemos que si la radiación se libera el año próximo, no tendremos ninguna defensa adecuada. Podemos creer que, a lo largo de los cuatro siglos próximos, se inventará algún tipo de contramedida, o alguna cura. De este modo podemos pensar que, si hay un escape radiactivo dentro de 400 años, entonces será mucho menos probable que cause muertes. Si somos *muy* optimistas, podemos creer que esto es mil millones de veces menos probable. Esto sería una razón diferente para descontar, por un factor de mil millones, muertes dentro de 400 años. No estaríamos haciendo la disparatada afirmación de que, si provocamos esas muertes, cada una de ellas importará moralmente mil millones de veces menos que una sola muerte el año próximo. En vez de esto estaríamos afirmando que estas muertes más remotas son mil millones de veces menos probables. Esta sería la razón por la cual, en nuestra opinión, apenas tenemos necesidad de preocuparnos por el escape radiactivo de dentro de 400 años. Si tuviéramos razón en afirmar que tales muertes son mil millones de veces menos probables, nuestra conclusión estaría justificada. Las muertes que no ocurren, tanto ahora como dentro de 400 años, no importan.

Este ejemplo ilustra un punto general. Debemos descontar las predicciones que tengan más probabilidad de ser falsas. Llamemos a esto la Tasa de Descuento Probabilístico. Las predicciones acerca del futuro más distante son falsas con mayor probabilidad. De forma que las dos clases de tasa de descuento, la *temporal* y la *probabilística*, correlacionan aproximadamente. Pero son muy diferentes. Por eso es un error descontar por el tiempo más bien que por la probabilidad. Una objeción es que esto declara erróneamente nuestra concepción moral. Nos hace decir, no que las malas consecuencias más remotas son menos probables, sino que son menos importantes. Esta no es nuestra opinión real. Una objeción mayor es que las dos tasas de descuento no siempre coinciden. Las predicciones sobre el futuro distante no son verdaderas con menor probabilidad a una tasa de *n* por ciento por año. Cuando las aplicábamos al futuro más distante, era *más* probable que muchas predicciones fuesen ciertamente verdaderas. Si descontamos por el tiempo más bien que por la probabilidad, puede que así seamos llevados a lo que, incluso según nuestras propias asunciones, son las conclusiones incorrectas.

> *El Argumento de los Costes de Oportunidad*: Es a veces mejor recibir un beneficio antes, puesto que este beneficio puede ser entonces usado para producir más beneficios. Si una inversión rinde una ganancia al año siguiente, esto valdrá más que la misma ganancia dentro de diez años, si la ganancia anterior puede ser reinvertida provechosamente a lo largo de estos diez años. Cuando hemos añadido los beneficios extra que proceden de esta reinversión, la suma total de beneficios será mayor. Un argumento similar incluye ciertas clases de coste. El retrasar algunos beneficios conlleva de este modo *costes de oportunidad*, y viceversa.

A veces se piensa que esto justifica una Tasa de Descuento Social. Pero la justificación fracasa, y por las mismas dos razones. Ciertos costes de oportunidad aumentan a través del tiempo. Pero si descontamos por el tiempo, más bien que limitarnos a añadir estos costes extra, representaremos mal nuestro razonamiento moral. Y lo que es más importante, podemos equivocarnos.

Consideremos esos beneficios que no se reinvierten sino que se consumen. Cuando se reciben más tarde, puede que esto no implique costes de oportunidad. Supongamos que vamos a decidir si construimos un aeropuerto. Esto destruiría una extensión de hermoso campo. Perderíamos el beneficio de disfrutar de esta belleza natural. Si no construimos el aeropuerto planeado, nosotros y nuestros sucesores disfrutaríamos de este beneficio cada año futuro. De acuerdo con una Tasa de Descuento Social, los beneficios en los años posteriores cuentan por mucho menos que el beneficio del año próximo. ¿Cómo podría justificar esto una apelación a los costes de oportunidad? El beneficio recibido el año siguiente —nuestro disfrute de esta belleza natural— no puede ser reinvertido con provecho.

Tampoco puede aplicarse el argumento a esos costes que son meramente «consumidos». Supongamos que sabemos que, si adoptamos determinada política, habrá un riesgo de causar deformidades genéticas. El argumento no puede demostrar que una deformidad genética el año próximo deba contar diez veces más que una deformidad dentro de veinte años. Como mucho lo que se podría decir es esto. Podríamos decidir que, por cada niño así afectado, la gran suma de *k* dólares proporcionaría una compensación adecuada. Si fuéramos a proporcionar tal compensación, el coste presente de asegurar esta sería mucho mayor para una deformidad causada el año próximo. Ahora tendríamos que poner aparte casi la totalidad de los *k* dólares. Un simple décimo de esta suma, si lo pusiera aparte ahora y lo invirtiera provechosamente, podría rendir en veinte años lo que entonces sería equivalente a *k* dólares. Esto aporta una razón para estar menos preocupado ahora por las deformidades que podríamos causar en el futuro distante. Pero la razón *no* es que tales deformidades importen menos. La razón es que ahora nos costaría sólo un décimo de esa cantidad asegurar que, cuando se produjesen tales deformidades, podríamos proporcionar compensación. Esta es una diferencia crucial. Supongamos que sabemos que de hecho no proporcionaremos compensación. Esto podría ocurrir, por ejemplo, si no pudiéramos identificar las particulares deformidades genéticas que nuestra política hubiera causado. Esto elimina nuestra razón para estar menos preocupados ahora por las deformidades de años posteriores. Si no vamos a pagar compensación alguna por tales deformidades, se convierte en un hecho irrelevante que, en el caso de deformidades posteriores, *habría* sido más barato asegurar ahora que *pudiéramos haber* pagado una compensación. Pero si este hecho nos ha llevado a adoptar una Tasa de Descuento Social, podemos fracasar a la hora de constatar cuándo se convierte en irrelevante. Podemos ser llevados a asumir que, aunque no haya compensación, las deformidades dentro de veinte años importan sólo un décimo de lo que importan las deformidades del año próximo.

Aquí tenemos otra objeción a este argumento. En ciertos períodos, la inversión *no* reporta ningún rendimiento. Cuando esto ocurre, el Argumento de los Costes de Oportunidad no puede aplicarse. La Tasa de Descuento Social fracasa una vez más en correlacionar con algo que importa.

Estos breves comentarios pasan por alto muchas de las cuestiones. De los diversos argumentos a favor de una Tasa de Descuento Social, la apelación a los Costes de Oportunidad es el más difícil de valorar. Pero la cuestión central es simple, me parece a mí. Puede ser en varios sentidos más conveniente, o más elegante, calcular los costes de oportunidad empleando una Tasa de Descuento Social. Pero las conclusiones que se establecen por tales cálculos podrían volverse a expresar de un modo temporalmente neutro. Al describir los efectos de futuras políticas, los economistas podrían determinar qué beneficios y costes futuros habría en momentos diferentes, de una manera que no empleara ninguna tasa de descuento. Los argumentos que apelan a los costes de oportunidad podrían exponerse del todo en estos términos. Me parece que, sobre cualquier cuestión importante que necesitemos decidir, esta sería una mejor descripción de las alternativas, puesto que es menos engañosa. Haría más fácil llegar a la decisión correcta.

El Argumento de que Nuestros Sucesores Estarán Mejor. Si suponemos que nuestros sucesores estarán mejor de lo que nosotros estamos ahora, hay dos argumentos plausibles para descontar los beneficios y los costes que les damos y les imponemos. Si medimos los beneficios y los costes en términos monetarios, ajustándolos a la inflación futura, podemos recurrir a la utilidad marginal decreciente del dinero. El mismo incremento de riqueza por lo general trae un beneficio más pequeño a los que están en mejor situación. Podemos recurrir también a un principio distributivo. Un beneficio igualmente grande dado a los que están mejor puede decirse que es menos importante moralmente.

Estos dos argumentos, aunque buenos, no justifican una Tasa de Descuento Social. La razón para descontar estos beneficios futuros no es que vayan más lejos en el futuro, sino que irán a personas más favorecidas. Aquí como en todas partes deberíamos decir lo que queremos decir. Y la correlación es otra vez imperfecta. Algunos de nuestros sucesores puede que no estén mejor de lo que estamos

nosotros ahora. Si no lo están, los argumentos que acabamos de dar no consiguen aplicarse.

> *El Argumento del Sacrificio Excesivo*. Una típica afirmación reza: «Necesitamos sin duda una Tasa de Descuento por razones teóricas. Porque si no, cualquier pequeño aumento de beneficios que se prolongue mucho en el futuro podría requerir cualquier cantidad de sacrificio en el presente, puesto que con el tiempo los beneficios compensarían el coste».

Se aplican las mismas objeciones. Si es por eso por lo que adoptamos una Tasa de Descuento Social, estaremos expresando erróneamente lo que creemos. Nuestra creencia no es que la importancia de los beneficios futuros se reduzca a un ritmo constante. Es más bien que no se puede exigir moralmente a ninguna generación que haga más que ciertas clases de sacrificios por el bien de las generaciones futuras. Si esto es lo que pensamos, esto es lo que debería influir en nuestras decisiones.

Podemos tener otra creencia. Si tenemos como meta la mayor suma neta de beneficios a través del tiempo, esto podría requerir una distribución desigual entre diferentes generaciones. Los utilitaristas dirían que, dados unos supuestos realistas, esto no sería verdadero. Pero supongamos que lo es. Entonces podemos desear evitar la conclusión de que debe darse tal distribución desigual. Y podemos evitarla, en algunos casos, si descontamos beneficios posteriores. Pero, como señala Rawls, este es el modo malo de evitar esta conclusión. Si no creemos que deba darse tal desigualdad, no deberíamos simplemente aspirar a la mayor suma neta de beneficios. Deberíamos tener un segundo fin moral: que los beneficios se compartan equitativamente entre las diferentes generaciones. A nuestro principio de utilidad deberíamos añadir un principio referente a la justa distribución. Esto expresa más fielmente nuestra concepción real. Y elimina la razón que teníamos para descontar beneficios posteriores.

Si en cambio expresamos nuestra concepción adoptando una Tasa de Descuento Social, podemos ser conducidos sin necesidad a conclusiones inverosímiles. Supongamos que, al *mismo* coste para

nosotros mismos ahora, podríamos prevenir o una catástrofe menor en el futuro más próximo o una mayor en el futuro más lejano. Como prevenir la catástrofe mayor no implicaría ningún coste extra, el Argumento del Sacrificio Excesivo no logra aplicarse. Pero si adoptamos el argumento para justificar una Tasa de Descuento, seremos llevados a concluir que vale menos la pena prevenir la catástrofe más grande.

> *El Argumento de las Relaciones Especiales*: Hay utilitaristas que afirman que cada persona debería dar un peso igual a los intereses de todo el mundo. No es lo que la mayor parte de nosotros piensa. De acuerdo con la Moralidad del Sentido Común, debemos dar algún peso a los intereses de los desconocidos. Pero hay determinadas personas a las que o podemos o debemos dar cierta clase de prioridad. Así, moralmente se nos permite dar cierta clase de prioridad a nuestros propios intereses. La mayoría de nosotros piensa que no tenemos deber alguno de ayudar a los demás si esto requiere de nosotros un sacrificio demasiado grande. Y hay ciertas personas a cuyos intereses *debemos* dar cierta clase de prioridad. Son aquellas con las que estamos en ciertas relaciones especiales. Por ejemplo, cada persona debe dar cierta clase de prioridad a los intereses de sus hijos, padres, alumnos, pacientes, aquellos a quienes representa, o sus conciudadanos.

Una concepción semejante se aplica naturalmente a los efectos de nuestros actos sobre las generaciones futuras. Nuestros sucesores inmediatos serán nuestros propios hijos. De acuerdo con el sentido común, debemos dar a su bienestar un peso especial. Podemos pensar lo mismo, aunque en un grado reducido, sobre los hijos de nuestros hijos.

Afirmaciones similares parecen plausibles a nivel comunitario. Creemos que nuestro gobierno debe estar especialmente preocupado por los intereses de sus propios ciudadanos. Sería natural decir que debe estar especialmente preocupado por los futuros hijos de sus ciudadanos, y, en un grado menor, por sus nietos.

Tales declaraciones podrían dar apoyo a una nueva clase de Tasa de Descuento. Aquí estaríamos descontando, no por el tiempo

mismo, sino por grados de parentesco. Pero al menos estas dos relaciones no pueden divergir radicalmente. Nuestros nietos no pueden nacer todos antes de todos nuestros hijos. Como la correlación es, aquí, más segura, podríamos tener la tentación de emplear una Tasa de Descuento estándar. Pienso que, también aquí, esto estaría injustificado. En primer lugar, según cualquier Tasa de Descuento los efectos más remotos siempre cuentan menos. Pero una Tasa de Descuento con respecto al Parentesco debería dejar de aplicarse en algún punto —o bien, para evitar la discontinuidad, aproximarse asintóticamente a un nivel horizontal que esté sobre cero—. Debemos dar *algún* peso a los efectos de nuestros actos sobre los simples desconocidos. No debemos dar *menos* peso a los efectos sobre nuestros propios descendientes.

Ni tampoco debería esta Tasa de Descuento aplicarse a toda clase de efectos. Consideremos esta comparación. Tal vez el gobierno de los Estados Unidos deba dar en general prioridad al bienestar de sus propios ciudadanos. Pero esto no se aplica a la imposición de daños graves. Supongamos que este gobierno decide reanudar las pruebas nucleares en la atmósfera. Si pronostica que las explosiones resultantes causarían varias muertes, ¿debería descontar las muertes de extranjeros? ¿Debería en consecuencia trasladar las pruebas al Océano Índico? Pienso que en tal caso, las relaciones especiales no representan ninguna diferencia moral. Deberíamos adoptar la misma concepción en relación con los daños que les imponemos a nuestros sucesores remotos.

He discutido seis argumentos a favor de la Tasa de Descuento Social. Ninguno tiene éxito. Lo más que podrían justificar es el uso de semejante tasa como tosca regla general. Pero esta regla a menudo se equivoca. A menudo puede ser moralmente permisible estar menos preocupado por los efectos más remotos de nuestras políticas sociales. Pero esto nunca sería *porque* estos efectos sean más remotos. Más bien sería porque es menos probable que ocurran, o serían efectos en personas que están mejor que nosotros, o porque sería ahora más barato asegurar la compensación, o sería por una de las otras razones que he dado. Todas estas diferentes razones tienen que exponerse y juzgarse por separado, según sus propios méritos.

Si las juntamos en un montón como Tasa de Descuento Social, nos hacemos a nosotros mismos moralmente ciegos.

La lejanía en el tiempo correlaciona más o menos con toda una gama de hechos moralmente importantes. También lo hace la lejanía en el espacio. Esos con quienes tenemos las mayores obligaciones, nuestra propia familia, con frecuencia viven con nosotros en la misma casa. Con frecuencia vivimos cerca de aquellos hacia quienes tenemos obligaciones especiales, nuestros clientes, alumnos, pacientes. La mayoría de nuestros conciudadanos viven más cerca de nosotros que la mayoría de los extranjeros. Pero nadie sugiere que, puesto que se dan estas correlaciones, deberíamos adoptar una Tasa de Descuento Espacial. Nadie piensa que estaríamos moralmente justificados si nos preocupáramos menos por los efectos de largo alcance de nuestros actos, a una tasa de *n* por ciento por metro. La Tasa de Descuento Temporal, me parece a mí, está igual de injustificada.

Cuando los otros argumentos no son de aplicación, debemos estar preocupados por igual por los efectos predecibles de nuestros actos tanto si estos van a ocurrir dentro de uno, de cien o de mil años. Esto tiene gran importancia. Algunos efectos son previsibles aun en el futuro distante. Los desechos nucleares pueden ser peligrosos durante miles de años. Y algunos de nuestros actos tienen efectos permanentes. Es lo que ocurriría por ejemplo con la destrucción de una especie, o de gran parte de nuestro medio, o de partes irreemplazables de nuestro patrimonio cultural.

G. Si hacer que alguien exista puede beneficiarle

Esta cuestión, curiosamente, ha sido descuidada. Así, en un informe de la Comisión del Senado de los Estados Unidos para el Crecimiento Demográfico y la Economía Americana, se afirma que «no habría beneficios sustanciales derivados del crecimiento continuo de la población de los Estados Unidos». Este informe no considera nunca si, en caso de que nacieran más americanos, esto podría beneficiar a estos americanos.

Si vamos a defender alguna concepción sobre la superpoblación, tenemos que considerar esta cuestión. Si un acto es una parte necesaria de la causa de la existencia de una persona con una vida digna de vivirse, ¿beneficia con ello este acto a esta persona? Defenderé que la respuesta Sí no está, como algunos afirman, evidentemente equivocada.

Algunos objetores afirman que la vida no se puede juzgar que sea ni mejor ni peor que la no existencia. Pero la vida de una cierta clase puede juzgarse que es o buena o mala —o digna de vivirse o digna de no vivirse—. Si una cierta clase de vida es buena, es mejor que nada. Si es mala, es peor que nada. Al juzgar que la vida de una persona es digna de vivirse, o mejor que nada, *no* tenemos por qué implicar que habría sido peor para esta persona que no hubiera existido nunca.

Los juicios de esta clase con frecuencia se realizan sobre la última parte de una vida. Consideremos a alguien muriendo penosamente que ya haya hecho sus despedidas. Esta persona puede decidir que seguir viva sería peor que morir. Para hacer este juicio, no tiene por qué *comparar* cómo habría sido seguir viviendo con *cómo habría sido haber muerto*. Como escribe Williams, «podría considerar lo que hay delante de ella, y decidir si quería o no quería pasar por ello» [16]. Y podría decidir, de manera parecida, que estaba contenta o por el contrario lamentaba lo que tenía por detrás. Podría decidir que, en algún punto del pasado, si hubiera sabido lo que tenía por delante, habría o no habría querido vivir el resto de su vida. Podría de este modo concluir que estas partes de su vida fueron mejores o peores que nada. Si semejantes afirmaciones pueden aplicarse a las partes de una vida, pueden aplicarse, me parece, a vidas enteras [17].

[16] Williams (2), pp. 85-6.

[17] En Williams (2), p. 87, también escribe Williams: «Nada de esto —incluyendo los pensamientos del suicida calculador— requiere mi reflexión sobre un mundo en que yo nunca existo en absoluto. En los términos de los "mundos posibles"... un hombre podría, según esta explicación, tener una razón desde su propio punto de vista para preferir un mundo posible en que él continuara existiendo por más tiempo a uno en el que siguiera existiendo menos tiempo, o al revés —como en el caso del suicida—. Pero no tendría ninguna razón de esta clase para preferir un mundo en que él no existiese en absoluto. El pensamiento sobre su ausencia total del mundo tendría que ser de una clase diferente, reflexiones impersonales sobre el valor para el mundo de su presencia o ausencia... Mientras que puede pensar de modo egoísta acerca de lo que sería para él vivir más o menos tiempo, no puede pensar de modo egoísta acerca de lo que sería para él no haber existido nunca en absoluto». Williams ha subrayado que, si alguien está pensando en el suicidio, no tiene por qué comparar lo que se extiende ante él con cómo sería estar muerto. Esta persona, simplemente, puede decidir si desea o no desea experimentar lo que se extiende ante ella. Esto puede tomarse como siendo la decisión de que, para ella, esta parte de su vida es mejor o peor que nada. En el fragmento recién citado, Williams sugiere que no se pueden tomar decisiones y hacer juicios semejantes sobre la totalidad de la propia vida. La razón apuntada es que alguien «no puede pensar de modo egoísta acerca de lo que sería para él no haber existido nunca». Alguien puede comprender claramente la posibilidad de que él podría no haber existido nunca. Si hay algo que no puede imaginar, sólo puede ser *como qué* sería no haber existido nunca, qué se sentiría entonces. Pero si alguien decide que su vida ha valido la pena, o desea no haber nacido nunca, no tiene por qué hacer esta comparación —justo como el hombre que piensa en el suicidio no necesita comparar el resto de su vida con cómo sería estar muerto.

Los que se opongan a esto podrían apelar ahora a

> *El Requisito de los Dos Estados*: Beneficiamos a alguien sólo si le hacemos estar mejor de lo que habría estado si no en ese momento.

Podrían decir: «Al hacer que alguien exista, y tenga una vida digna de vivirse, no estamos haciendo que esta persona esté *mejor* de lo que habría estado si no hubiéramos hecho nada. Esta persona no habría estado *peor* si no hubiera existido nunca».

Para evaluar este argumento, primero deberíamos hacer la pregunta siguiente. Si alguien existe ahora, y tiene una vida que vale la pena vivirse, ¿está mejor de lo que estaría ahora si hubiera muerto y hubiera dejado de existir? Supongamos que respondemos que Sí. Al aplicar el Requisito de los Dos Estados, consideramos el haber dejado de existir como un estado en el que alguien puede estar peor. ¿Por qué no podemos decir lo mismo sobre no existir nunca? ¿Por qué no podemos afirmar que, si alguien existe ahora con una vida que vale la pena vivirse, está mejor de lo que estaría si nunca exis-

tiera? Es verdad que *no existir nunca* no es un estado corriente. Pero tampoco lo es *haber dejado de existir*. ¿Dónde está nuestro error si los tratamos igual cuando aplicamos el Requisito de los Dos Estados?

Podría contestarse que cuando muere alguien hay una persona concreta que ha dejado de existir. Podemos referirnos a ella. En contraste, no hay personas concretas que no existan nunca. No podemos referirnos a ninguna de tales personas.

Esto podría ser una buena contestación si estuviéramos afirmando que, al hacer que las personas no existan nunca, podríamos estarles perjudicando. Pero estamos haciendo una afirmación diferente. Esta es la de que, al causar que alguien exista, podemos estarle beneficiando. Como esta persona *sí* que existe, podemos referirnos a ella cuando describimos la alternativa. Sabemos quién es el que, en esta posible alternativa, nunca habría existido. En los casos que estamos considerando, no se da la supuesta diferencia entre haber dejado de existir y no existir nunca. Igual que podemos referirnos a la persona que ahora podría haber dejado de existir, podemos referirnos a la persona que podría no haber existido. No se nos ha mostrado por qué, al aplicar el Requisito de los Dos Estados, no deberíamos tratar a estos dos estados de la misma manera.

El defensor del Requisito de los Dos Estados podría a renglón seguido cambiar de opinión en lo que se refiere al estado de estar muerto, o haber dejado de existir. Podría afirmar que no es este un estado en el que alguien pueda estar peor. Entonces podría afirmar lo mismo sobre no existir nunca.

Con esta revisión, el Requisito de los Dos Estados se vuelve demasiado fuerte. Implica que salvarle la vida a alguien no puede beneficiarle, puesto que la persona salvada no está en mejor situación de lo que lo habría estado si hubiera dejado de existir. En el caso de salvar la vida a alguien, ahora sería defendible relajar el Requisito de los Dos Estados. Entendemos la razón específica por la que, en este caso, el Requisito no se satisface. Podemos afirmar que, a causa de este rasgo especial del caso, el Requisito no necesita satisfacerse aquí. Si el resto de la vida de alguien fuese digno de vivirse, podemos considerar que salvar su vida es un caso especial de beneficiarle. Y si podemos relajar el Requisito en el caso de salvarle la vida, ¿por qué

no podemos hacer lo mismo en el caso de darle la vida? Si la vida de alguien es digna de vivirse, ¿por qué no podemos considerar que causar que viva es un caso especial de beneficiarle?

Los opositores podrían volverse ahora a

> *El Requisito Comparativo Pleno*: Beneficiamos a alguien sólo si hacemos lo que será mejor para él.

Ellos podrían decir: «Al causar que alguien exista, no podemos estar haciendo lo que será mejor para él. Si no hubiésemos causado que exista, no habría sido peor para él». A diferencia de la última forma del Requisito de los Dos Estados, este nuevo requisito permite que salvarle la vida a alguien pueda beneficiarle. Podemos afirmar que morir puede ser peor para alguien, aunque esto no le haga sentirse peor. (Estaríamos rechazando aquí la *Concepción Lucreciana* de que un suceso puede ser malo para alguien sólo si le hace sufrir posteriormente, o al menos lamentarse de él.)

Puesto que incluye salvar vidas, el Requisito Comparativo Pleno es más plausible que la forma más fuerte del Requisito de los Dos Estados. Pero si podemos relajar el último en nuestros dos casos especiales, puede ser defendible relajar el primero en el caso de darle la vida a alguien. Podemos admitir que, en toda otra clase de caso, beneficiamos a alguien sólo si hacemos lo que será mejor para él [18]. En el caso de darle a alguien la vida, comprendemos la razón especial por la que la alternativa no habría sido peor para él. Podríamos decir que, en este caso especial, el Requisito no necesita ser satisfecho. Supongamos que hemos concedido que salvarle la vida a alguien puede beneficiarle. Si mi propia vida es digna de vivirse, entonces me habría beneficiado haber tenido mi vida a salvo en todo momento después de que comenzara. ¿Tengo que afirmar que, mientras que me benefició haber tenido mi vida a salvo *sólo después* de que comenzara, no me benefició haberla tenido comenzada? Puedo negar esta afirmación de una manera defendible.

[18] Esto no es verdadero según nuestro uso corriente de «beneficio». Pero, como sostengo en la Sección 25, sí que lo es según el uso moralmente significativo.

Hacer que alguien exista es un caso especial porque la alternativa no habría sido peor para esta persona. Podemos admitir que, por esta razón, hacer que alguien exista no puede ser *mejor* para esta persona. Pero puede ser *bueno* para ella [19]. En este paso de «mejor» a «bueno», admitimos que el Requisito Comparativo Pleno no se satisface. Pero aún podríamos hacer dos clases de comparaciones. Si puede ser bueno para alguien hacerle vivir, *hasta qué punto* es esto bueno para esta persona dependerá de lo buena que sea su vida —de hasta qué punto su vida es digna de vivirse—. Y podemos hacer comparaciones interpersonales. Supongamos que la vida de *Jack* es digna de vivirse, pero no por un amplio margen. Si sus ataques de depresión se hiciesen más frecuentes y más severos, él empezaría a dudar de que valiera la pena seguir viviendo su vida. En contraste, la vida de *Jill* es perfectamente digna de vivirse. Entonces podemos afirmar que, cuando causamos que Jack existiera, esto fue bueno para Jack, pero fue mucho *menos* bueno para Jack de lo que fue para Jill hacer que existiera Jill.

Estas afirmaciones evitan una objeción común. Cuando afirmamos que fue bueno para alguien que se le hiciera existir, no estamos implicando que, si no se le hubiese hecho existir, esto habría sido malo para él. Y nuestras declaraciones se aplican sólo a personas que son o serían reales. No hacemos afirmaciones sobre personas que permanecerían siempre meramente posibles. No estamos diciendo que sea malo para personas posibles el que no lleguen a ser reales.

Finalizo con estas observaciones. He considerado tres cosas: no existir nunca, comenzar a existir y dejar de existir. He sugerido que, de las tres, comenzar a existir debería clasificarse con dejar de existir. A diferencia de no existir nunca, comenzar a existir y dejar de existir son cosas que les ocurren a las personas reales. Por eso podemos afirmar que pueden ser buenas o malas para estas personas. La afirmación contraria es que comenzar a existir debería ser clasificada con no existir nunca, y que ninguna de las dos cosas puede ser buena o mala para las personas. La razón que a veces se da es que, si *no* hubiéramos comenzado a existir, nunca habríamos existido, lo

[19] Debo esta sugerencia, y mucho más en la Cuarta Parte, a J. McMahan.

cual no habría sido malo para nosotros. Pero no estamos diciendo que comenzar a existir pueda ser bueno o malo para las personas *cuando no ocurre*. Nuestra afirmación se refiere a comenzar a existir *cuando ocurre*. Admitimos una diferencia entre comenzar a existir y dejar de existir. Para casi todos los sucesos, si su ocurrencia fuese buena para las personas, su no ocurrencia habría sido peor para las mismas. Pero, podemos sugerir, hay un suceso especial cuya no ocurrencia no habría sido peor para esta persona real. Este suceso, como era de esperar, es el llegar a ser real de esta persona.

Estas observaciones no son concluyentes. Se podrían presentar más objeciones. Lo que afirmo es sólo que, si creemos que causar que se exista puede beneficiar, esta creencia es defendible. He apelado a tres puntos. Primero, no necesitamos afirmar que sea malo para personas posibles que nunca lleguen a ser reales. Como dice Nagel: «Todos nosotros...somos afortunados por haber nacido. Pero... no puede decirse que no haber nacido sea una desgracia» [20]. Segundo, si me benefició haber tenido mi vida a salvo sólo después de que comenzara, no estoy obligado a negar que me benefició haberla tenido comenzada. Desde mi punto de vista presente, no hay ninguna distinción profunda entre estas dos cosas. (Podría negarse que me beneficiara haber tenido mi vida a salvo. Pero si se afirma esto, se convierte en irrelevante la cuestión de si hacer que alguien exista puede beneficiarle. Debo salvarle la vida a un niño que se está ahogando. Si con ello no beneficio a este niño, esta parte de la moralidad no puede explicarse en términos que tengan en cuenta a las personas afectadas.) Tercero, causar que alguien exista es evidentemente un caso especial. Hay quienes argumentan que esto no es un beneficio porque carece de un rasgo que comparten todos los otros beneficios. Pero este argumento da por sentado lo que hay que demostrar. Como se trata de un caso especial, puede ser una excepción a alguna regla general. Apelar a una regla general simplemente supone que no puede haber excepciones.

Ha habido un debate similar acerca de si *existente* es un *predicado* o una propiedad genuina que podrían poseer los objetos. Hay quie-

[20] Nagel (4), p. 7.

nes afirman que, como carece de algunos de los rasgos de otros predicados, *existente* no es un predicado. Otros afirman que esto sólo demuestra que *existente* es un predicado *peculiar*. Podemos decir, de forma parecida, que causar que alguien exista, alguien que tendrá una vida digna de vivirse, le da a esta persona un beneficio especial.

H. Principios rawlsianos

Muchos aplican a ejemplos concretos los tipos de principio que Rawls emplea sólo dentro del contexto de su más amplia teoría. Por eso digo que estos principios no son de Rawls, sino solamente *rawlsianos*. (Marx se quejaba de que alguno de sus seguidores era *plus marxiste que moi*. Rawls podría quejarse de lo mismo.)

Consideremos cualquier caso en que, en los diferentes resultados, existirían las mismas personas. De acuerdo con Maximin, el mejor resultado de todos es aquel en el que los menos favorecidos resultaran más favorecidos. Podemos pensar que este resultado tiene que ser lo que es lo mejor de todo para los que están peor. Pero esto no es así.

Supongamos que soy médico. Dependiendo de lo que haga, las consecuencias podrían ser:

(1) *Jack, Bill y John* quedan completamente paralíticos.
(2) Jack y John se curan. Bill queda completamente paralítico.
(3) Jack y Bill se curan. John queda parcialmente paralítico.

John sería más difícil de curar que Bill, quien sería más difícil de curar que Jack. Por eso no puedo hacer más para curarlos que lo que hago en los resultados (2) y (3). Si curo a John no puedo curar a Bill, y si curo a Bill sólo puedo curar a John parcialmente.

(3) es el resultado en que la persona que peor está resulta más favorecida. Pero este no es el resultado que es el mejor para esta persona. (3) es peor para ella de lo que habría sido (2). Maximin selecciona de manera correcta (3) como el resultado mejor. Pero no podemos decir que, si seguimos Maximin, esto tiene que ser lo

mejor para los que están en peor situación. Seguir Maximin puede ser peor para estas personas que algo distinto que pudiéramos haber hecho. Cuando nos damos cuenta de este detalle, todavía podemos aceptar Maximin. Pero tenemos que revisar otros dos Principios Rawlsianos.

Consideremos primero el *Principio de Diferencia*. En el resultado (3) hay desigualdad. Según el Principio de Diferencia, causar una desigualdad tal es injusto a no ser que produzca el resultado que sea el mejor de todos para los que están en la peor situación. En el resultado (3) es John el que está en la peor situación. Al producir esta desigualdad no he producido el resultado que es el mejor de todos para John, dado que (2) habría sido mejor para él. Según el Principio de Diferencia, la desigualdad en el resultado (3) es injusta. Por una razón parecida, la desigualdad en el resultado (2) es injusta. Para evitar la injusticia, tengo que actuar de una de dos maneras. O bien no tengo que curar a nadie, o tengo que curar parcialmente a todas las tres personas. Aunque podría curar del todo a dos de ellas y parcialmente a la tercera, no debo hacerlo. Esta es evidentemente la conclusión equivocada. Si deseamos aplicarlo a ejemplos especiales, tenemos que revisar el Principio de Diferencia para que deje de implicar conclusiones semejantes.

No necesito discutir esta revisión, puesto que el Principio de Diferencia no se aplica a mis resultados imaginarios A y A +. Este principio se aplica sólo a la desigualdad que es tanto creada deliberadamente como evitable. La desigualdad en A + no es de esta clase.

Consideremos a continuación lo que llamo el *Principio de Selección*. Este afirma que el mejor resultado de todos es aquel que es el mejor para los que están en la peor situación. En mi ejemplo médico, (3) no puede ser el mejor puesto que es peor que (2) para la persona que en (3) está peor. Esta es evidentemente la conclusión equivocada. Tenemos que revisar el Principio de Selección para que deje de implicar conclusiones semejantes.

La revisión obvia es Maximin. Este difiere del Principio de Selección precisamente de la manera que necesitamos para implicar, correctamente, que el mejor de todos los resultados es (3).

El resto de este Apéndice lo escribió John Broome, después de que hubiéramos discutido las cuestiones tratadas arriba.

En Rawls, uno de los principios de justicia es el Principio de Diferencia, que Rawls especifica así (p. 302):

> «Las desigualdades económicas y sociales deben ser organizadas de manera que redunden... en el mayor beneficio de los menos favorecidos».

Esta formulación no expresa exactamente lo que Rawls pretendía con ella. Supongamos que la India en 1800 pudiese haber tenido una cualquiera de tres constituciones. Cada una habría distribuido equitativamente otros bienes primarios, pero habrían distribuido el bienestar económico y social de la manera siguiente:

> Según la Constitución (1) los hindúes y los británicos reciben 100 ambos.
> Según la Constitución (2) los hindúes reciben 120 y los británicos 110.
> Según la Constitución (3) los hindúes reciben 115 y los británicos 140.

Es evidente que Rawls piensa en el Principio de Diferencia para favorecer la Constitución (3). En ocasiones (p. e., p. 152) describe el Principio de Diferencia como una «regla maximin», y (3) satisface esta regla. Pero no satisface la formulación que cité arriba. Los menos favorecidos bajo la Constitución (3) son los hindúes, y las mayores desigualdades bajo (3), comparadas con (2), no redundan en beneficio de los hindúes. Habrían salido mejor parados bajo (2).

Desde luego, se podría cambiar fácilmente la redacción para expresar lo que quiere decir Rawls con más fidelidad. Pero varios argumentos de Rawls en pro del Principio de Diferencia dependen efectivamente de su propia redacción. Esto no ocurre con su argumento principal, la afirmación de que las personas en la Posición Original seleccionaría la regla maximin, pero muchos de sus otros argumentos dan por sentado que el Principio de Diferencia favorece arreglos sociales que redundan en el mayor beneficio de las personas menos favorecidas. Estos argumentos fracasan en ejemplos como el mío. Por ejemplo, en la p. 103 Rawls dice: «"B" —el hombre representativo de los menos favorecidos— puede aceptar que A esté en mejor situación puesto que los privilegios de A se han ganado de maneras que promocionan las posibilidades de "B"». No es cierto que los privilegios de la persona británica representativa bajo la Constitución (3) se hayan ganado de maneras que promocionen las posibilidades del hindú representativo.

Podría ser que Rawls no estuviera satisfecho con mi modo de tratar el ejemplo. Organizaciones sociales alternativas, dice él (pp. 95-100), deben compararse en términos de las situaciones de los grupos, o de los hombres representativos que representan a los grupos, más bien que de los individuos. Y un grupo debe definirse simplemente por su nivel de renta y de riqueza. (Pero ver p. 29, donde Rawls admite que los grupos puedan a veces identificarse por la raza.) De forma que lo que debemos comparar no es la posición de los hindúes bajo las Constituciones (2) y (3), sino la posición del grupo menos favorecido, que está formado por diferentes personas en los dos casos. El grupo menos privilegiado bajo (3) está en mejor situación que el grupo menos privilegiado bajo (2). Esta es la comparación que a Rawls le podría gustar que hiciésemos. Pero mi cuestión es que sus *argumentos* a menudo asumen implícitamente que estamos comparando posiciones alternativas que el mismo *individuo* podría ocupar. En el argumento que cité, Rawls compara las posiciones alternativas del hombre representativo de los menos favorecidos, pero no hay ningún hombre que represente a los grupos menos favorecidos tanto bajo (2) como bajo (3). Sólo podemos asumir que el hombre que representa a los menos favorecidos es el hindú representativo, y como digo el argumento no funciona con él.

Es evidente que para justificar la elección de la Constitución (3) antes que la (2) tenemos que comparar los intereses de dos grupos. Tenemos que decir que la pérdida de los hindúes al escoger (3) es menor que la pérdida de los británicos al escoger (2). De forma que tenemos que considerar los intereses de diferentes personas en la misma medida en que lo hacen los utilitaristas, y en la

misma medida en que se suponía que el principio de diferencia lo evita (pp. 175-83). Si (3) realmente es mejor que (2), yo debería decir que sólo puede ser sobre bases parecidas a las utilitaristas. Y si las bases utilitaristas van por el otro camino —si, digamos, hay muchos más hindúes que británicos— yo debería encontrar muy difícil de creer que (3) es realmente mejor.

John Broome. (Me gustaría agradecer a John Rawls sus comentarios.)

I. Lo que hace que la vida de alguien vaya mejor

¿Qué sería lo mejor para alguien? ¿Qué iría más a favor de sus intereses? ¿Qué haría que la vida de esta persona marchara, para ella, lo mejor posible? A las respuestas a estas preguntas las llamo *teorías del propio interés*. Hay tres clases de teorías. Para las *Teorías Hedonistas*, aquello que sería lo mejor para alguien es lo que hiciese su vida lo más feliz posible. Para las *Teorías de la Realización de Deseos*, aquello que sería lo mejor para alguien es lo que, a lo largo de su vida, realizase sus deseos de la mejor manera posible. Para las *Teorías de la Lista*
832 *Objetiva*, ciertas cosas son buenas o malas para nosotros, independientemente de que queramos tener las buenas o evitar las malas.

Los *hedonistas estrictos* suponen, falsamente, que el placer y el dolor son dos clases distintivas de experiencia. Compárese los placeres de satisfacer la sed o el deseo sexual intensos, de escuchar música, resolver un problema intelectual, leer una tragedia, y saber que nuestro hijo es feliz. Estas diversas experiencias no contienen ninguna cualidad común distintiva.

Lo que los dolores y los placeres tienen en común son sus relaciones con nuestros deseos. Según el uso de «dolor» que tiene significación racional y moral, todos los dolores son indeseados cuando los experimentamos, y un dolor es peor o mayor cuanto menos lo deseemos. De manera similar, todos los placeres son deseados cuando los experimentamos, y son mejores o mayores cuanto más los deseemos. Estas son las afirmaciones del *Hedonismo de la Preferencia*. Para esta concepción, una de dos experiencias es más placentera si es preferida.

Esta teoría no necesita seguir los usos corrientes de las palabras «dolor» y «placer». Supongamos que yo pudiera irme a una fiesta a disfrutar de los diversos placeres de comer, beber, reírme, bailar y hablar con mis amigos. En vez de ello me podría quedar en casa y leer *El Rey Lear*. Sabiendo cómo serían las dos alternativas, prefiero leer *El Rey Lear*. Amplía el uso corriente decir que esto me daría más placer. Pero según el Hedonismo de la Preferencia, si añadimos algunas asunciones suplementarias dadas abajo, leer *El Rey Lear* me haría pasar una tarde mejor. Griffin cita un caso más extremo. Próximo al final de su vida, Freud rechazó tomar drogas para quitarse el dolor, prefiriendo pensar atormentado que estar confusamente eufórico. De estos dos estados mentales, la euforia es más agradable. Pero según el Hedonismo de la Preferencia, pensar atormentado era para Freud un estado mental mejor. Resulta más claro aquí no extender el significado de la palabra «agradable». Un hedonista de la preferencia debería afirmar simplemente que, puesto que Freud prefirió pensar con claridad aunque sufriendo un tormento, su vida fue mejor si fue como él la prefirió [21].

Consideremos a continuación las Teorías de la Realización de Deseos. La más simple es la Teoría *No Restringida*. Afirma que lo que es lo mejor para alguien es aquello que realizaría del mejor modo 833
posible *todos* sus deseos, a lo largo de su vida. Supongamos que me encuentro con un desconocido que tiene lo que se piensa es una enfermedad mortal. Se despierta mi simpatía, y llego a desear intensamente que el desconocido se cure. Nunca nos volvemos a encontrar. Más tarde, sin que yo lo sepa, el extranjero se cura. Según la Teoría No Restringida de la Realización de Deseos, este suceso es bueno para mí, y hace que mi vida vaya mejor. Pero esto no es plausible. Deberíamos rechazar esta teoría.

Otra teoría apela sólo a nuestros deseos sobre nuestra propia vida. La llamo la *Teoría del Éxito*. Difiere del Hedonismo de la Preferencia sólo en un aspecto. La Teoría del Éxito apela a *la totalidad* de nuestras preferencias acerca de nuestra propia vida. Un hedonista de la preferencia apela sólo a preferencias acerca de esos ras-

[21] Griffin (1).

gos de nuestra vida que son discernibles de manera introspectiva. Supongamos que deseo intensamente que los demás no me engañen. Según el Hedonismo de la Preferencia será mejor para mí si creo que no me están engañando. Será irrelevante si mi creencia es falsa, puesto que esto no representa ninguna diferencia para mi estado mental. Según la Teoría del Éxito, será peor para mí que mi creencia sea falsa. Tengo un intenso deseo en relación con mi propia vida —que no debería ser engañado así—. Es malo para mí que este deseo no se realice, aunque yo crea falsamente que se realiza.

Cuando esta teoría apela solamente a deseos que versan sobre nuestra propia vida, puede no estar claro lo que esto excluye. Supongamos que quiero que mi vida sea tal que todos mis deseos, sean cuales sean sus objetos, se realicen. Puede parecer que esto hace coincidir a la Teoría del Éxito, cuando se me aplicase a mí, con la Teoría No Restringida de la Realización de Deseos. Pero un teórico del éxito debe afirmar que este deseo no versa en realidad sobre mi propia vida. Esto es como la distinción entre un cambio real en algún objeto, y el así denominado *cambio Cambridge.* Un objeto sufre un cambio Cambridge si hay algún cambio en las afirmaciones verdaderas que pueden hacerse sobre él. Supongamos que me corto la cara afeitándome. Esto causa en mí un cambio real. También causa un cambio en Confucio. Llega a ser verdadero, de Confucio, que él vivió en un planeta en el que después se corto una cara más. Este es meramente un cambio Cambridge.

Supongamos que soy un exiliado y no puedo ponerme en contacto con mis hijos. Quiero que sus vidas vayan bien. Yo podría afirmar que quiero vivir la vida de alguien con hijos cuyas vidas vayan bien. Un teórico del éxito debe afirmar de nuevo que este no es en realidad un deseo sobre mi propia vida. Si sin yo saberlo una avalancha mata a uno de mis hijos, esto no es malo para mí y no hace que mi vida marche peor.

Un teórico del éxito *tendría* en cuenta algunos deseos parecidos. Supongamos que trato de darles a mis hijos un buen comienzo en la vida. Trato de darles la educación correcta, buenos hábitos y fortaleza psicológica. Una vez más, ahora soy un exiliado, y nunca podré enterarme de lo que les ocurre a mis hijos. Supongamos que,

sin que yo lo sepa, la vida de mis hijos va mal. Uno se encuentra con que la educación que yo le di le hace inútil para el trabajo, otro tiene una crisis psiquiátrica y el otro se convierte en un ladrón de poca monta. Si la vida de mis hijos fracasa de estas formas, y si estos fracasos son en parte resultado de los errores que cometí en calidad de padre, estos fracasos de las vidas de mis hijos serían juzgados según la Teoría del Éxito como malos para mí. Uno de mis deseos más intensos era tener éxito como padre. Lo que les está ocurriendo ahora a mis hijos, aunque yo no lo sepa, muestra que este deseo no se ha realizado. Mi vida fracasó en uno de los aspectos en que yo más deseaba que tuviera éxito. Aunque ignoro este hecho, es malo para mí, y hace verdadero que yo he tenido una vida peor. Es como el caso en que deseo intensamente que no me engañen. Aunque nunca lo sepa, es malo para mí tanto que me engañen como si resulto ser un mal padre. No son estas diferencias introspectivamente discernibles en mi vida consciente; de manera que, para el Hedonismo de la Preferencia, estos sucesos no son malos para mí. Pero según la Teoría del Éxito sí que lo son.

Consideremos a continuación los deseos que algunos tienen sobre lo que suceda después de muertos. Para un hedonista de la preferencia, una vez que estoy muerto, nada malo puede sucederme. Un teórico del éxito debería negarlo. Volvamos al caso en que todos mis hijos llevan vidas desgraciadas a causa de los errores que cometí en calidad de padre. Supongamos que las vidas de mis hijos van todas mal sólo después de que yo he muerto. Resulta que mi vida ha sido un fracaso, en uno de los aspectos que más me importaban. Un teórico del éxito debería afirmar que, también aquí, esto hace verdadero que yo tuve una vida peor.

Hay teóricos del éxito que rechazarían esta afirmación, desde el momento en que nos dicen que ignoremos los deseos de los muertos. Pero supongamos que me preguntan, «¿Quieres que sea verdadero, aun después de que estés muerto, que fuiste un buen padre?». Yo respondería «Sí». En relación con mi deseo es irrelevante que se realice antes o después de que yo muera. Estos teóricos del éxito consideran que es malo para mí que mis intentos fracasen, aunque, porque soy un exiliado, nunca lo vaya a saber. ¿Cómo puede impor-

tar entonces, cuando mis intentos fracasan, que esté muerto? Todo lo que mi muerte hace es *asegurar* que nunca lo sabré. Si pensamos que es irrelevante que nunca vaya a saber de la no realización de mis deseos, no podemos afirmar justificablemente que mi muerte represente una diferencia.

Vuelvo ahora a preguntas y objeciones que surgen tanto del Hedonismo de la Preferencia como de la Teoría del Éxito.

¿Deberíamos apelar sólo a los deseos y las preferencias que alguien tiene realmente? Volvamos a mi elección entre ir a una fiesta o quedarme en casa leyendo *El Rey Lear*. Supongamos que, sabiendo cómo serían ambas alternativas, elijo quedarme en casa. Y supongamos que después nunca me arrepiento de esta elección. Según una teoría, esto pone de manifiesto que quedarme en casa a leer *El Rey Lear* me proporcionó una mejor tarde. Pero esto es un error. Podría ser cierto que, si yo hubiese elegido ir a la fiesta, nunca habría lamentado esta elección. De acuerdo con esta teoría, esto habría demostrado que ir a la fiesta me proporcionó una mejor tarde. Esta teoría implica, así, que cada alternativa habría sido mejor que la otra. Como implica semejantes contradicciones, la teoría tiene que revisarse. La revisión obvia es no apelar sólo a mis preferencias reales, en la alternativa que elijo, sino también a las preferencias que habría tenido si hubiera elegido de otro modo [22].

En este ejemplo cualquier alternativa que elija nunca lo lamentaría. Si esto es así, ¿podemos afirmar aún que una de las alternativas me daría una mejor tarde? Según algunas teorías, cuando en dos alternativas yo tuviera tales preferencias opuestas, ninguna alternativa sería mejor o peor para mí. Esto no es plausible cuando una de mis preferencias opuestas hubiera sido mucho más fuerte. Supongamos que, si decido ir a la fiesta, voy a estar sólo ligeramente contento de haber hecho esta elección, pero que si decido quedarme a leer *El Rey Lear* voy a estar muy contento. Si esto es así, leer *El Rey Lear* me da una mejor tarde.

[22] Véase «Prudence» [«Prudencia»], por P. Bricker, *The Journal of Philosophy*, julio 1980.

Independientemente de que apelemos al Hedonismo de la Preferencia o a la Teoría del Éxito, no deberíamos apelar sólo a los deseos o las preferencias que tengo realmente, sino también a los que yo habría tenido en las diversas alternativas que estuvieron abiertas para mí en diferentes momentos. Una de estas alternativas sería la mejor para mí si fuera aquella en la que yo hubiera realizado los deseos y las preferencias más intensos. Esto nos permite afirmar que alguna vida alternativa habría sido mejor para mí, aunque a lo largo de mi vida real yo esté contento de haberla elegido a ella en lugar de su alternativa.

Hay otra distinción que se aplica tanto al Hedonismo de la Preferencia como a la Teoría del Éxito. Estas teorías son *sumativas* si apelan a todos los deseos de una persona, reales e hipotéticos, que versan o bien sobre sus estados mentales o bien sobre su vida. Al decidir cuál alternativa produciría la mayor suma neta total de deseos realizados, asignamos un número positivo a cada deseo que se realice, y un número negativo a cada deseo que no se realice. Cuán grandes sean estos números depende de la intensidad de los deseos en cuestión. (En el caso de la Teoría del Éxito, que apela a deseos pasados, puede depender también de cuánto duran estos deseos. Como sugiero en el capítulo 8, esto puede ser una debilidad en esta teoría. El problema no surge para el Hedonismo de la Preferencia, que sólo apela a los deseos que tenemos, en momentos diferentes, en relación con nuestros estados mentales presentes.) La suma neta total de deseos realizados es la suma de los números positivos menos los números negativos. Contando con que podemos comparar la fuerza relativa de los diferentes deseos, en teoría se podría realizar este cálculo. La elección de una unidad para los números no afecta al resultado.

Otra versión de ambas teorías no apela, de esta manera, a todos los deseos y las preferencias que tiene una persona en relación con su propia vida. Apela sólo a deseos y preferencias *globales* más bien que *locales*. Una preferencia es global si versa sobre una parte de la vida de uno considerada como una totalidad, o trata acerca de la vida completa de uno. Las versiones *globales* de estas teorías considero que son más plausibles.

Consideremos este ejemplo. Sabiendo que tú aceptas una teoría sumativa, te digo que estoy a punto de hacer que tu vida marche mejor. Te inyectaré una droga que crea adicción. De ahora en adelante, te despertarás cada mañana con un deseo sumamente intenso de ponerte otra inyección de esta droga. Tener este deseo no será en sí mismo ni agradable ni doloroso, pero si el deseo no se satisface en una hora, entonces se hará muy doloroso. Pero esto no debe preocuparte, porque te daré abundantes cantidades de la droga en cuestión. Cada mañana podrás satisfacer este deseo inmediatamente. Ni la inyección ni sus efectos secundarios serían tampoco agradables ni dolorosos. Te pasarás el resto de tu vida como estás ahora.

¿Qué implicarían las Teorías Sumativas en relación con este caso? Podemos suponer justificadamente que no aceptarías mi propuesta. Preferirías no hacerte adicto a la droga, aunque te asegure que nunca te va a faltar. Podemos suponer también con causa justificada que, si yo sigo adelante, tú lamentarás siempre haberte convertido en un adicto a esta droga. Pero es probable que tu deseo inicial de no convertirte en un adicto, y tus remordimientos posteriores por haberlo hecho, no serían tan fuertes como los deseos que tienes cada mañana de otra inyección. Dados los hechos tal y como los he descrito, tu razón para preferir no convertirte en un adicto no sería muy fuerte. Te podría disgustar la idea de ser adicto a algo, y lamentarías la incomodidad menor que supondría tener que acordarte siempre de llevar contigo suficientes provisiones de droga. Pero estos deseos podrían ser mucho más débiles que los que tendrías cada mañana de una inyección recién preparada.

Según las Teorías Sumativas, si hago de ti un adicto estaré incrementando el total de tus deseos realizados. Estaré haciendo que uno de tus deseos no se realice, el de no convertirte en un adicto, el cual, tras mi acto, se convierte en el deseo de curarte. Pero también te estaré proporcionando una serie indefinida de deseos sumamente fuertes, uno cada mañana, todos los cuales los puedes satisfacer. La realización de todos estos deseos compensaría la no realización de tus deseos de no llegar a ser un adicto y de curarte. Según las Teorías Sumativas, al hacer de ti un adicto te estaré beneficiando —estaré haciendo que tu vida vaya mejor.

Esta conclusión no es admisible. Tener estos deseos, y tenerlos satisfechos, ni es agradable ni es doloroso. No hace falta ser hedonistas para creer, con más justificación, que de ninguna manera es mejor para ti tener y realizar esta serie de intensos deseos.

¿Podrían revisarse las Teorías Sumativas para hacer frente a esta objeción? ¿Hay algún rasgo de los deseos adictivos que justificaría la afirmación de que deberíamos ignorarlos cuando calculamos el total de tus deseos realizados? Podríamos decir que se pueden ignorar porque son deseos que preferirías no tener. Pero esta no es una revisión aceptable. Supongamos que tienes un gran dolor. Tienes ahora un deseo muy intenso de no estar en el estado en el que estás. Según tu teoría revisada, un deseo no cuenta si prefieres no tenerlo. Esto tiene que aplicarse a tu intenso deseo de no estar en el estado en el que estás. Preferirías no tener este deseo. Si no te disgustara el estado en el que estás, no sería doloroso. Como nuestra teoría revisada no cuenta los deseos que preferirías no tener, implica, de manera absurda, que no puede ser malo para ti tener un gran dolor.

Puede haber otras revisiones que podrían hacer frente a estas objeciones. Pero resulta más sencillo apelar a las versiones globales del Hedonismo de la Preferencia y de la Teoría del Éxito. Estas apelan sólo a los deseos que tiene alguien en relación con una parte de su vida considerada como un todo, o en relación con su vida entera. Las Teorías Globales nos dan la respuesta correcta en el caso en que yo hago de ti un adicto. Tú preferirías no convertirte en un adicto, y después preferirías dejar de serlo. Estas son las únicas preferencias a las que las Teorías Globales apelan: ignoran tus deseos concretos de una inyección recién preparada cada mañana, puesto que tú ya has considerado estos deseos al formar tu preferencia global.

Este caso imaginario de la adicción es, en sus rasgos esenciales, similar a un sinnúmero de otros casos. Hay innumerables casos en que es verdadero tanto (1) que, si la vida de una persona marchara de una de dos maneras, esto produciría una suma total mayor local de deseos realizados, pero (2) que la otra alternativa es la que ella globalmente preferiría, *cualquiera que fuera* el modo en que su vida real marchara.

En lugar de describir otro de los innumerables casos reales, mencionaré un caso imaginario. Este es el análogo, dentro de una vida, de la *Conclusión Repugnante* que discuto en la Cuarta Parte. Supongamos que yo podría o bien tener cincuenta años de vida de una calidad sumamente alta, o un número indefinido de años que apenas vale la pena vivir. En la primera alternativa, según cualquier teoría, mis cincuenta años irían sumamente bien. Yo sería muy feliz, lograría grandes cosas, haría mucho bien, y amaría y sería amado por muchas personas. En la segunda alternativa mi vida siempre sería digna de vivirse, aunque no por mucho. No habría nada malo en esta vida, y cada día contendría unos pocos pequeños placeres.

Según las Teorías Sumativas, si la segunda vida fuese lo suficientemente larga sería mejor para mí. Cada día de esta vida tengo algunos deseos sobre mi vida que se realizan. En los cincuenta años de la primera alternativa, habría una suma local muy grande de deseos realizados. Pero esta sería una suma finita, y finalmente sería superada por la suma de deseos realizados en mi indefinidamente larga segunda alternativa. Una forma más simple de poner este punto sería esta. La primera alternativa sería buena. En la segunda, como mi vida es digna de vivirse, vivir cada día extra es bueno para mí. Si meramente sumamos lo que es bueno para mí, algún número de estos días extra produciría la mayor suma total.

Yo no creo que la segunda alternativa me diera una vida mejor. Por eso rechazo las Teorías Sumativas. Es probable que, en ambas alternativas, yo globalmente prefiriera la primera. Como las Teorías Globales implicarían entonces que la primera alternativa me da una vida mejor, son estas teorías las que me parecen más plausibles [23].

Volvámonos ahora al tercer tipo de teoría que mencioné: la Teoría de la Lista Objetiva. De acuerdo con ella, hay ciertas cosas que son buenas o malas para las personas, independientemente de que las personas quieran tener las cosas buenas o evitar las cosas malas. Las cosas buenas podrían incluir bondad moral, actividad

[23] Véase otra vez Bricker.

racional, el desarrollo de las capacidades propias, tener hijos y ser una buena madre o un buen padre, el conocimiento y el disfrute de la verdadera belleza. Las cosas malas podrían incluir ser traicionado, manipulado, calumniado, engañado, privado de libertad o de dignidad, y gozar con placer sádico o con placer estético de lo que de hecho es feo [24].

Un teórico de la lista objetiva podría afirmar que su teoría coincide con la versión global de la Teoría del Éxito. Según esta teoría, lo que haría que mi vida marchara mejor depende de lo que yo prefiriera, ahora y en las diversas alternativas, si conociera todos los hechos relevantes acerca de estas alternativas. Un teórico de la lista objetiva podría decir que los hechos más relevantes de todos serían los que acabamos de mencionar —los hechos acerca de lo que sería bueno o malo para mí—. Y podría afirmar que cualquiera que conociera estos hechos desearía lo que es bueno para él, y querría evitar lo que es malo.

Aunque esto fuera verdadero, por mucho que la Teoría de la Lista Objetiva coincidiera con la Teoría del Éxito, las dos teorías seguirían siendo distintas. Un teórico del éxito rechazaría esta descripción de la coincidencia. Según su teoría, nada es bueno o malo para las personas *sean cuales sean* sus preferencias. Algo es malo para alguien sólo cuando, si conociera los hechos, desearía evitarlo. Y los hechos relevantes no incluyen los supuestos hechos citados por el teórico de la lista objetiva. Según la Teoría del Éxito es malo para una persona ser engañada, por ejemplo, en caso de que, y porque, no sea esto lo que ella quiere. El teórico de la lista objetiva hace la afirmación opuesta. Las personas no quieren ser engañadas porque esto es malo para ellas.

Como estos comentarios implican, hay una diferencia importante entre, por una parte, el Hedonismo de la Preferencia y la Teoría del Éxito, y por otra parte la Teoría de la Lista Objetiva. Las primeras dos clases de teorías dan una explicación del propio interés que es puramente descriptiva —que no apela a hechos acerca del valor—. Esta explicación apela sólo a lo que una prefiere y preferi-

[24] Véase, por ejemplo, Moore y Ross (2).

ría, supuesto un conocimiento completo de los hechos puramente no evaluativos sobre las alternativas. En contraste, la Teoría de la Lista Objetiva apela directamente a lo que afirma son hechos acerca del valor.

A la hora de elegir entre estas teorías, tenemos que decidir cuánto peso damos a casos imaginarios en que las preferencias plenamente informadas de una persona son estrafalarias. Si podemos apelar a estos casos, ponen en tela de juicio tanto el Hedonismo de la Preferencia como la Teoría del Éxito. Consideremos al hombre que imaginó Rawls, ese que quiere pasarse la vida contando las briznas de hierba de diferentes praderas. Supongamos que este hombre sabe que podría conseguir grandes progresos si en lugar de ello se pusiera a trabajar en una parte especialmente útil de la Matemática Aplicada. Aunque podría lograr tan significativos resultados, prefiere seguir contando briznas de hierba. Según la Teoría del Éxito, si dejamos que cubra todos los casos imaginables, podría ser mejor para esta persona contar sus briznas de hierba en lugar de lograr resultados matemáticos importantes y útiles.

El contraejemplo podría ser más ofensivo. Supongamos que lo que alguien preferiría en mayor medida, conociendo las alternativas, fuese una vida en la que, sin ser descubierto, causara tanto dolor como pudiese a los demás. Para la Teoría del Éxito, una vida así sería lo mejor para esta persona.

Puede que seamos incapaces de aceptar estas conclusiones. ¿Debemos por ello abandonar esta teoría? Esto es lo que hizo Sidgwick, aunque los que le citan raramente lo notan. Él sugiere que «el bien futuro de un hombre, en general, es lo que desearía y buscaría ahora en general, si todas las consecuencias de todas las diferentes líneas de conducta que le están abiertas fuesen certeramente previstas y adecuadamente realizadas en la imaginación en el momento temporal presente» [25]. Como comenta: «La noción de "Bien" así alcanzada tiene un elemento ideal: es algo que no *es* siempre realmente deseado por los seres humanos, no es algo a lo que siempre aspiren realmente: pero el elemento ideal es enteramente

[25] Sidgwick (1), pp. 111-12.

interpretable en términos *de hecho*, real o hipotético, y no introduce ningún juicio de valor». Sidgwick rechaza entonces esta explicación, afirmando que lo que es en último término bueno para alguien es lo que esta persona *desearía* si sus deseos estuvieran en armonía con la razón. Se necesita esta última frase, pensaba Sidgwick, para excluir los casos en que los deseos de la persona son irracionales. Él da por sentado que hay cosas que tenemos buenas razones para desear, y otras que tenemos buenas razones para no desear. Podrían ser las cosas de las que sostienen las Teorías de la Lista Objetiva que son buenas o malas para nosotros.

Supongamos que estuviéramos de acuerdo en que, en algunos casos imaginarios, lo que alguien desearía en mayor medida tanto ahora como después, sabiendo todo lo que hay que saber sobre las alternativas, *no* fuese lo que sería lo mejor para él. Si aceptamos esta conclusión, puede parecer que tenemos que rechazar tanto el Hedonismo de la Preferencia como la Teoría del Éxito. Tal vez, como Sidgwick, tengamos que poner restricciones a lo que puede desearse racionalmente.

Podría afirmarse en cambio que podemos descartar la apelación a tales casos imaginarios. Podría afirmarse que lo que las personas preferirían de hecho, si conocieran los hechos relevantes, sería siempre algo que podríamos aceptar como lo que es realmente bueno para ellas. ¿Es esta una buena respuesta? Si estamos de acuerdo en que en los casos imaginarios lo que alguien preferiría podría ser algo que es malo para él, en estos casos hemos abandonado nuestra teoría. Si esto es así, ¿podemos defender nuestra teoría diciendo que, en los casos reales, no estaría errada? Creo que no es esta una defensa adecuada. Pero no seguiré aquí con esta cuestión.

Esta objeción puede aplicarse con menos fuerza al Hedonismo de la Preferencia. Según esta teoría, lo que puede ser bueno o malo para alguien sólo pueden ser rasgos discernibles de su vida consciente. Son los rasgos que, en el momento, él desea o no desea. Pregunté arriba si es malo para las personas ser engañadas porque prefieren no serlo o si prefieren no ser engañadas porque es malo para ellas. Consideremos la cuestión paralela con respecto al dolor. Hay quienes han afirmado que el dolor es intrínsecamente malo, y

que por eso nos provoca aversión. Como he dado a entender, pongo en duda esta afirmación. Después de tomar ciertas clases de drogas, la gente dice que la cualidad de sus sensaciones no se ha alterado, pero que ya no les provocan aversión esas sensaciones. Consideraríamos estas drogas como analgésicos eficaces. Esto sugiere que la maldad de un dolor consiste en que provoca aversión, y que no provoca aversión porque sea malo. El desacuerdo entre estas concepciones necesitaría mucha más discusión. Pero, si la segunda concepción es mejor, es más verosímil afirmar que sea lo que sea lo que alguien quiere o no quiere experimentar —por muy estrafalarios que encontremos sus deseos— deberían tenerse en cuenta como siendo para esta persona verdaderamente agradables o dolorosos, y como siendo por esa razón buenos o malos para ella. (Puede haber todavía casos en que sea plausible afirmar que sería malo para alguien que disfrutara de ciertas clases de experiencias; se podría decir, por ejemplo, del placer sádico. Pero puede que haya pocos de tales casos.)

Si en cambio apelamos a la Teoría del Éxito, no nos importará sólo la cualidad experimentada de nuestra vida consciente. Nos importarán cosas tales como si estamos logrando lo que tratamos de lograr, si nos están engañando, etc. Al considerar esta teoría, podemos afirmar verosímilmente más a menudo que, aunque la persona conociera los hechos, sus preferencias podrían equivocarse, y fracasar a la hora de corresponder con lo que sería bueno o malo para ella.

¿Cuál de estas diferentes teorías deberíamos aceptar? No intentaré dar aquí una respuesta. Pero voy a terminar mencionando otra teoría, que podría decirse que combina lo que resulta más convincente de estas teorías rivales. Llama la atención que los que han tratado esta cuestión hayan discrepado tan radicalmente. Muchos filósofos han sido hedonistas convencidos; muchos otros han sostenido con igual convencimiento que el Hedonismo es un gran error.

Hay hedonistas que han llegado a su concepción como sigue. Consideran una posición contraria, como la que afirma que lo que es bueno para alguien es tener conocimiento, tomar parte en la actividad racional y ser conscientes de la verdadera belleza. Estos hedonistas preguntan, «¿Serían buenos estos estados mentales si no trajesen consigo ningún placer, y si la persona que se encontrara en ellos no tuviera el más mínimo deseo de que continuaran?». Puesto que responden No, llegan a la conclusión de que el valor de estos estados mentales tiene que radicar en su ser gustados, y en el hecho de que hacen surgir el deseo de que continúen.

Este razonamiento da por sentado que el valor de un todo no es más que la suma del valor de sus partes. Si eliminamos la parte a la que apela el hedonista, lo que queda parece no tener valor, por tanto, el Hedonismo es la verdad.

Supongamos en cambio, lo que es más verosímil, que el valor de un todo no puede ser la mera suma del valor de sus partes. Podríamos decir entonces que lo que es lo mejor para las personas es un compuesto. No es sólo que se hallen en los estados conscientes en los que quieren estar. Tampoco es sólo que tengan conocimiento, emprendan la actividad racional o sean conscientes de la verdadera belleza, y cosas por el estilo. Lo que es bueno para una persona no es sólo lo que dicen los hedonistas, ni sólo lo que dicen los teóricos de la lista objetiva. Podríamos pensar que si tuviéramos *una* de estas cosas *sin la otra*, lo que tuviéramos tendría poco valor o no tendría ninguno. Podríamos afirmar, por ejemplo, que lo que es bueno o malo para alguien es tener conocimiento, hallarse entregado a la actividad racional, experimentar amor mutuo, y ser consciente de la belleza, al mismo tiempo que se desean intensamente justo estas cosas. Según esta concepción, cada bando en esta disputa vio sólo media verdad. Cada uno presentó como suficiente algo que era sólo necesario. El placer con muchas otras clases de objetos no tiene valor. Y, si están completamente desprovistos de placer, no hay valor en el conocimiento ni en la actividad racional ni en el amor ni en la conciencia de la belleza. Lo que tiene valor, o es bueno para alguien, es tener los dos: entregarse a estas actividades y desear intensamente estar así entregado [26].

[26] Véase Edwards, de principio a fin. Una sugerencia semejante la hace Platón en el *Filebo*. Para una discusión más profunda de las diferentes teorías del propio interés, véase J. Griffin, *Well-Being* [*El bienestar*] (OUP, 1986).

J. La concepción de Buda

Al comienzo de su conversación, el rey, cortésmente, le pregunta al monje su nombre, y recibe la siguiente respuesta: «Señor, se me conoce como "Nagasena": mis compañeros en la vida religiosa se dirigen a mí como "Nagasena". Aunque mis padres (me) pusieron el nombre "Nagasena"... es sólo una denominación, una forma de hablar, una descripción, una usanza convencional. "Nagasena" es sólo un nombre, porque ninguna persona se encuentra aquí» [27].

> Un ser sensible sí que existe, ¿así lo crees, Oh Mara?
> Estas engañado por una concepción falsa.
> Este manojo de elementos está vacío de Yo,
> En él no hay ningún ser sensible.
> Igual que una colección de partes de madera,
> Recibe el nombre de carro,
> También nosotros le damos a unos elementos
> El nombre de un ser ficticio [28].

Buda ha hablado así: «¡Oh hermanos!, las acciones sí que exis-
846 ten, y también sus consecuencias, pero la persona que actúa no. No hay nadie que tire esta colección de elementos, y nadie que asuma una colección nueva de ellos. No existe ningún Individuo, es sólo un nombre convencional dado a una colección de elementos [29]».

Vasubandhu: ...Cuando Buda dice, «Yo mismo fui este maestro Sunetra», quiere decir que su pasado y su presente pertenecen al mismo linaje de existencias momentáneas; no quiere decir que los elementos primeros no desaparecieran. Igual que cuando decimos «este mismo fuego que ha sido visto consumiendo esa cosa ha alcanzado este objeto», el fuego no es el mismo, pero pasando por alto esta diferencia llamamos indirectamente fuego a la continuidad de sus momentos [30].

Vatsiputriya: Si no hay Alma, ¿quién es el que recuerda? *Vasubandhu*: ¿Cuál es el significado de la palabra «recordar»? *Vatsiputriya*: Significa asir un objeto con la memoria. *Vasubandhu*: ¿Este «asir con la memoria» es algo diferente de la memoria? *Vatsiputriya*. Es un agente que actúa a través de la memoria. *Vasubandhu*: Ya hemos explicado la capacidad de actuar por la cual se produce la memoria. La causa productora de un recuerdo es un adecuado estado mental, nada más. *Vatsiputriya*: Pero cuando usamos la expresión «Caitra recuerda», ¿qué es lo que significa? *Vasubandhu*: En la corriente de fenómenos que se designa con el nombre *Caitra*, aparece un recuerdo [31].

El término budista para el individuo, un término que pretende indicar la diferencia entre la concepción budista y otras teorías, es *santana*, p. e., una «corriente» [32].

Vatsiputriya: ¿Qué es una esencia real y qué una nominal? *Vasubandhu*: Si existe algo por sí mismo (como un elemento separa-
do), tiene una existencia real. Pero si algo representa una combina- 847
ción (de tales elementos) es una existencia nominal [33].

> Lo mental y lo material están realmente aquí,
> Pero aquí no hay ningún ser humano que se pueda encontrar.
> Porque está vacío y simplemente fabricado como una muñeca,
> Nada más que sufrimiento apilado como hierba y leña [34].

[27] Del *Milina Panha*, citado en Collins, pp. 182-3.
[28] *Cila Mara*, citado en Th. Stcherbatsky, «The Soul Theory of the Buddhists» [«La teoría del alma de los budistas»], *Bulletin de l'Academie des Sciences de Russie*, 1919, p. 839.
[29] *Vasubandhu*, citado en Stcherbatsky, *op. cit.*, p. 845.
[30] Citado en Stcherbatsky, *op. cit.*, p. 851.
[31] *Op. cit.*, p. 853.
[32] Véase Collins, pp. 247-61.
[33] T. Stcherbatsky, *The Central Conception of Buddhism* [*La concepción central del Budismo*], Royal Asiatic Society, Londres, 1923, p. 26.
[34] El *Visuddhimagga*, citado en Collins, p. 133.

EPÍLOGO

La falta de importancia de la identidad* [1]

Podemos empezar con un poco de ciencia ficción. Aquí en la Tierra, entro en el teletransportador. Al apretar un botón, una máquina destruye mi cuerpo al mismo tiempo que graba los estados exactos de todas mis células. La información es enviada por radio a Marte, donde otra máquina hace una copia perfecta de mi cuerpo a partir de materiales orgánicos. La persona que despierta en Marte parece recordar haber vivido mi vida hasta el momento en que yo apreté el botón, y es en todos los demás aspectos exactamente igual que yo.

De los que han reflexionado sobre casos como este, algunos creen que sería yo el que despertara en Marte. Consideran el teletransporte simplemente la manera más rápida de viajar. Pero otros

* Este trabajo fue publicado en *Identity*, editado por Henrry Harris, OUP, 1995, y ha sido cedido amablemente por el autor para la presente edición española.

[1] Parte de esta conferencia hace uso de la Tercera Parte de mis *Reasons and Persons* [*Razones y Personas*], OUP, 1984. El material más nuevo será más completamente desarrollado en mi contribución a *Derek Parfit and His Critics* [*Derek Parfit y sus críticos*], editado por Jonathan Dancy, Blackwell, de próxima publicación.

creen que, si yo eligiera ser teletransportado, estaría cometiendo un terrible error. Según su opinión, la persona que despierta sería una mera réplica mía.

I

Este es un desacuerdo acerca de la identidad personal. Para comprender desacuerdos tales tenemos que distinguir dos clases de identidad. Dos bolas de billar pueden ser cualitativamente idénticas, o exactamente iguales. Pero no son numéricamente idénticas, o una y la misma bola. Si pinto de diferente color una de estas bolas, dejará de ser cualitativamente idéntica consigo misma como era, pero todavía será la misma bola. Consideremos a continuación una afirmación como esta: «Desde su accidente, ella ya no es la misma persona». Esto implica los dos sentidos de la identidad. Quiere decir que *ella*, la misma persona, *no* es ahora la misma persona. Esto no es una contradicción. Se afirma sólo que el carácter de esta persona ha cambiado. Esta persona numéricamente idéntica es ahora cualitativamente diferente.

Cuando los psicólogos discuten la identidad, por regla general lo que les interesa es la clase de persona que alguien es o quiere ser. Esta es la cuestión implicada, por ejemplo, en una crisis de identidad. Pero cuando los filósofos discuten la identidad, en lo que piensan es en la identidad numérica. Y en nuestra preocupación por nuestro propio futuro eso es lo que tenemos en mente. Tal vez crea que después de mi boda seré una persona diferente. Pero eso no hace de mi boda la muerte. Por mucho que yo cambie, aún estaré vivo si va a haber alguien con vida que será yo. De forma semejante, si yo fuera teletransportado mi Réplica en Marte sería cualitativamente idéntica a mí. Pero, a los ojos de los escépticos, no *sería* yo. *Yo* habré dejado de existir. Y damos por descontado naturalmente que es eso lo que importa.

Las preguntas acerca de nuestra identidad numérica adoptan todas ellas la forma siguiente. Tenemos dos modos de referirnos a una persona, y preguntamos si estos son modos de referirnos a la

misma persona. De esta manera podríamos preguntar si Boris Nikolayevich es Yeltsin. En las más importantes preguntas de este tipo, nuestros dos modos de referirnos a una persona seleccionan a una persona en momentos diferentes. De esta forma podríamos preguntar si la persona con la que hablamos ahora es la misma que la persona con la que ayer hablamos por teléfono. Estas son preguntas sobre la identidad a través del tiempo.

Para contestarlas, tenemos que conocer el *criterio* de la identidad personal: la relación entre una persona en un momento determinado y otra persona en otro momento distinto que hace que sean la misma persona.

Se han propuesto diferentes criterios. Según una concepción, lo que me hace el mismo a lo largo de toda mi vida es el hecho de que tengo el mismo cuerpo. Este criterio exige continuidad corporal ininterrumpida. Y no se da semejante continuidad entre mi cuerpo en la Tierra y el cuerpo de mi Réplica en Marte, de forma que, según este modo de ver el asunto, mi Réplica no sería yo. Otros autores apelan a la continuidad psicológica. Así, Locke afirmaba que si yo fuese consciente de una vida pasada en un cuerpo diferente, yo sería la persona que vivió esa vida. Para ciertas versiones de esta concepción, mi Réplica sería yo.

Los partidarios de estas diferentes opiniones con frecuencia recurren a casos en que entran en conflicto. La mayoría de ellos son puramente imaginarios, como el teletransporte. Hay filósofos que objetan que, como nuestro concepto de persona descansa en un andamiaje de hechos, no deberíamos esperar que tuviese aplicación en casos imaginarios en que hacemos abstracción de estos hechos. De acuerdo. Pero pienso que vale la pena tomar en consideración tales casos, por una razón diferente. Y es que podemos usarlos para descubrir, si no la verdad, sí lo que nosotros creemos. Podríamos haber encontrado que cuando consideramos casos de ciencia ficción nos limitamos a encoger los hombros. Pero no es así. Muchos de nosotros descubrimos que tenemos ciertas creencias sobre la clase de hecho que la identidad personal es.

Estas creencias salen a la luz del mejor modo cuando pensamos en tales casos desde el punto de vista de la primera persona. Así que cuando yo imagino que me está ocurriendo algo tú deberías imagi-

nar que te está ocurriendo a ti. Supongamos que vivo en un siglo futuro en el que la tecnología está avanzadísima, y que estoy a punto de sufrir una operación. Tal vez vayan a remodelar mi cuerpo y mi cerebro, o a reemplazarlos parcialmente. Habrá una persona resultante, que mañana se despertará. Pregunto: «¿Será yo esa persona? ¿O yo estoy a punto de morir? ¿Es este el final?». Tal vez no sepa responder a esta pregunta. Pero es natural asumir que tiene que *haber* una respuesta. Puede parecer que la persona resultante tiene que ser yo o alguien distinto. Y que la respuesta tiene que ser todo-o-nada. Esa persona no puede ser *en parte* yo. Si va a tener dolor mañana, este dolor no puede ser mío en parte. De manera que, podemos darlo por sentado, o bien sentiré ese dolor o bien no lo sentiré.

Si esto es lo que pensamos de casos así, estamos asumiendo que nuestra identidad tiene que ser *determinada*. Estamos asumiendo que en todo caso imaginable las preguntas acerca de nuestra identidad tienen que tener respuesta, y esa respuesta tiene que ser, sencillamente, Sí o No.

Preguntémonos ahora: ¿Puede ser verdad todo esto? Sólo hay una concepción desde la que podría serlo. Según ella, hay sustancias inmateriales: las almas o los Egos Cartesianos. Entidades estas que tienen las propiedades especiales que una vez se adscribieron a los átomos: son indivisibles, y su existencia continua es, en su misma naturaleza, todo o nada. Y un Ego como estos es lo que cada uno de nosotros realmente es.

A diferencia de varios autores, yo considero que esta concepción podría haber sido verdadera. Pero no contamos con buena evidencia para pensar que lo sea, y sí con alguna para pensar que no lo es. De manera que asumiré que ninguna concepción como esta es verdadera.

Si no creemos que haya Egos Cartesianos, ni otras entidades por el estilo, deberíamos aceptar el tipo de concepción que en otros lugares he denominado *Reduccionismo*. Según ella

> (1) La existencia de una persona sólo consiste en la existencia de un cuerpo y en la ocurrencia de una serie de pensamientos, experiencias y otros sucesos mentales y físicos.

Hay reduccionistas que afirman

> (2) Las personas no son otra cosa que sus cuerpos.

Esta idea puede que no parezca reduccionista, puesto que no reduce las personas a nada diferente. Pero ello es así sólo porque es hiperreduccionista: reduce las personas a los cuerpos de una manera tan enérgica que ni siquiera distingue entre ambos. Podemos denominarla Reduccionismo Identificador.

Una concepción como esta me parece demasiado simple. Pienso que deberíamos combinar (1) con

> (3) Una persona es una entidad que tiene un cuerpo, y que tiene pensamientos y otras experiencias.

Según esta concepción, aunque una persona es distinta de su cuerpo, y de cualquier serie de pensamientos y experiencias, la existencia de la persona nada más que *consiste* en ellos. De forma que podemos denominar a esta concepción Reduccionismo *Constitutivo*.

Puede ser de utilidad considerar otros ejemplos de esta clase de concepción. Si fundimos una estatua de bronce, la destruimos, pero no destruimos ese pedazo de bronce. De modo que aunque la estatua no consista nada más que en el pedazo de bronce, no son la misma cosa. De forma semejante, la existencia de una nación no consiste nada más que en la existencia de un grupo de personas en un territorio, que viven juntas de ciertas maneras. Pero la nación no es lo mismo que ese grupo de personas o ese territorio.

Consideremos a continuación el Reduccionismo Eliminativo. Semejante concepción es en ocasiones una respuesta a los argumentos en contra de la concepción identificadora. Supongamos que empezamos afirmando que una nación no es más que un grupo de personas en un territorio. Entonces nos persuadimos de que esto no puede ser, que el concepto de nación es el de una entidad que es distinta de sus personas y de su territorio. Podemos concluir que, en tal caso, realmente no existen cosas tales como naciones. Hay sólo grupos de personas, que viven juntas de ciertos modos.

En el caso de las personas, algunos textos budistas adoptan una concepción eliminadora. De acuerdo con ellos

(4) No hay en realidad cosas tales como las personas, sólo cerebros y cuerpos, pensamientos y otras experiencias.

Por ejemplo

Buda ha hablado así: «¡Oh hermanos!, las acciones sí que existen, y también sus consecuencias, pero la persona que actúa no... No existe ningún Individuo, es sólo un nombre convencional dado a una colección de elementos» [2].

O:

Lo mental y lo material están realmente aquí,
Pero aquí no hay ninguna persona que se pueda encontrar.
Porque está vacío y fabricado como una muñeca,
Nada más que sufrimiento apilado como hierba y leña [3].

El Reduccionismo Eliminativo a veces está justificado. Por ejemplo, tenemos razón cuando decimos que en realidad no hubo brujas sino sólo mujeres perseguidas. Pero el Reduccionismo respecto de un tipo de entidad con frecuencia no está bien expresado con la afirmación de que no hay tales entidades. Deberíamos admitir que hay naciones, y que nosotros, que somos personas, existimos.

En vez de afirmar que no hay entidades de un tipo, los reduccionistas deberían distinguir tipos de entidad o maneras de existir. Cuando la existencia de un X consiste sólo en la existencia de un Y, o de unos Ys, aunque el X sea *distinto* del Y o de los Ys, no es una entidad *independiente* o *que exista separadamente*. Las estatuas no existen separadamente de la materia de la que están hechas, ni tampoco existen las naciones separadamente de sus ciudadanos y de su territorio. De forma similar, creo que

[2] Vasubandhu, citado en Theodore Stcherbatsky, «The Soul Theory of the Buddhists» [«La teoría del alma de los budistas»], *Bulletin de l'Academie des Sciences de Russie*, 1919, p. 845.

[3] El *Visuddhimagga*, citado en Steven Collins, *Selfless Persons* [*Personas desinteresadas, sin yo*], Cambridge University Press, 1982.

(5) Aunque las personas son distintas de sus cuerpos y de cualquier serie de sucesos mentales, no son entidades independientes o que existan separadamente.

Los Egos Cartesianos, si existieran, no sólo serían distintos de los cuerpos humanos, sino que también serían entidades independientes. Se asevera que tales Egos son como los objetos físicos, salvo que son completamente mentales. Si hubiera tales entidades tendría sentido suponer que podrían dejar de estar causalmente relacionadas con un cuerpo y a pesar de ello seguir existiendo. Pero, según una concepción reduccionista, las personas no son independientes de sus cuerpos en ese sentido. (Lo que no es lo mismo que afirmar que nuestros pensamientos y demás experiencias sean meramente cambios en los estados de nuestros cerebros. Los reduccionistas, no creyendo en sustancias puramente mentales, pueden no obstante ser dualistas.)

Podemos volver ahora a la identidad personal a través del tiempo, o a qué constituye la existencia continua de la misma persona. Aquí una pregunta que se plantea es esta: ¿Qué explica la unidad de la vida mental de una persona? ¿Qué hace de pensamientos y experiencias tenidos en momentos diferentes los pensamientos y las experiencias de una única persona? De acuerdo con algunos no reduccionistas, esta pregunta no puede contestarse en otros términos. Simplemente tenemos que afirmar que estos pensamientos y experiencias diferentes son todos tenidos por la misma persona. Este hecho no consiste en ningunos otros hechos sino que es una verdad desnuda o última.

Si cada uno de nosotros fuese un Ego Cartesiano, así podría ser. Puesto que tal Ego sería una sustancia independiente, podría ser un hecho irreductible que diferentes experiencias fuesen todas cambios en los estados del mismo Ego persistente. Pero eso no podría ser verdadero de las personas, creo, si ellas no son entidades que existen separadamente, no obstante ser distintas de sus cuerpos. Así concebida, una persona no es la clase de entidad sobre la que podría haber verdades irreductibles como estas. Cuando las experiencias en diferentes momentos son todas tenidas por la misma persona, este hecho consiste en ciertos otros hechos.

Si no creemos en Egos Cartesianos, deberíamos afirmar

(6) La identidad personal a través del tiempo consiste sólo en continuidad física y/o psicológica.

A esta afirmación le podríamos dar contenido de diferentes maneras. Para una versión de esta concepción, lo que hace de experiencias diferentes las experiencias de una única persona es que son o bien cambios en los estados del mismo cerebro incorporado, o al menos relacionados con él de forma directa y causal. Esa tiene que ser la opinión de los que piensan que las personas son sólo cuerpos. Y podríamos mantener esa opinión aunque, como creo que deberíamos hacer, distingamos a las personas de sus cuerpos. Pero podríamos apelar, además o en lugar de ello, a varias relaciones psicológicas entre diferentes estados y sucesos mentales, tales como las que están implicadas en la memoria o en la persistencia de las intenciones, los deseos y otros elementos psicológicos. Eso es lo que quiero decir con continuidad psicológica.

Según el Reduccionismo Constitutivo, el hecho de la identidad personal es distinto de estos hechos acerca de la continuidad física y psicológica. Pero, como no consiste más que en ellos, no es un hecho independiente o que se dé separadamente. No es una diferencia más en lo que ocurre.

Para ilustrar esa distinción, consideremos un caso más sencillo. Supongamos que ya sé que hay varios árboles en una loma. Luego me entero de que, como esto es verdadero, hay un bosquecillo en esta loma. Esa no sería nueva información fáctica. Simplemente habría aprendido que un grupo de árboles como ese se puede denominar un «bosquecillo». La única información nueva que tengo es acerca de nuestro lenguaje. Que esos árboles puedan llamarse un bosquecillo no es, a no ser en un sentido trivial, un hecho sobre los árboles.

Algo parecido sucede en el caso más complicado de las naciones. Para conocer los hechos de la historia de una nación, basta con saber qué hicieron y dijeron grandes cantidades de personas. Los hechos sobre las naciones no pueden ser verdaderos de manera básica: tienen

que consistir en hechos sobre personas. Y una vez que conocemos estos otros hechos, todas las preguntas que queden sobre las naciones no son preguntas adicionales sobre lo que realmente ocurrió.

Creo que, del mismo modo, los hechos sobre las personas no pueden ser verdaderos de manera básica. Su verdad tiene que consistir en la verdad de los hechos acerca de cuerpos y acerca de varios sucesos físicos y mentales interrelacionados. Si conociéramos estos otros hechos, tendríamos toda la entrada empírica que nos hace falta. Si comprendiéramos el concepto de persona y no tuviéramos falsas creencias sobre lo que las personas son, sabríamos entonces, o bien lograríamos entender, la verdad de todas las afirmaciones adicionales sobre la existencia o la identidad de las personas. Sería así porque esas afirmaciones no nos dirían más sobre la realidad.

Este es el bosquejo más básico de la Concepción Reduccionista. Estas observaciones pueden tornarse más claras si nos volvemos a los denominados «casos problema» de la identidad personal. En un caso de estos nos imaginamos que sabemos que, entre yo ahora y una persona en el futuro, habrá ciertas clases o grados de continuidad o conexividad física y/o psicológica. Pero aunque conocemos estos hechos no podemos responder a la pregunta de si esa persona futura sería yo.

Como podemos ser de diferente opinión sobre cuáles son los casos problema, necesitamos más de un ejemplo. Consideremos primero la gama de casos que en otra parte he llamado el *Espectro Físico*. En cada uno de estos casos, alguna proporción de mi cuerpo sería reemplazada, en una operación única, por duplicados exactos de las células existentes. En el caso situado en el extremo cercano de esta gama, no sería reemplazada ninguna célula. En el caso situado en el extremo lejano, sería destruido y replicado todo mi cuerpo. Ese es el caso con el que comencé: el teletransporte.

Supongamos que creemos que en ese caso, en que todo mi cuerpo sería reemplazado, la persona resultante no sería yo sino una simple Réplica. Si no fuera reemplazada ninguna célula, la persona resultante sería yo. Pero ¿qué ocurre con los casos que hay entre medias, aquellos en los que el porcentaje de células reemplazado sería, digamos, el 30%, o el 50% o el 70%? ¿Aquí la persona resul-

tante sería yo? Cuando consideramos algunos de estos casos, no sabemos si responder Sí o No.

Supongamos a continuación que creemos que, incluso en el teletransporte, mi Réplica sería yo. Entonces deberíamos considerar una versión diferente de ese caso, aquella en que el escáner recibiría su información sin destruir mi cuerpo, y se fabricaría mi Réplica mientras yo todavía estoy vivo. En esta versión del caso, podemos admitir que mi Réplica no sería yo. Eso puede hacer vacilar nuestra convicción de que, en la versión original del caso, ella *sería* yo.

Si todavía mantenemos esa convicción, deberíamos volvernos a lo que he denominado el *Espectro Combinado*. En esta segunda gama de casos, se darían todos los diferentes grados tanto de conexividad física como de conexividad psicológica. Las nuevas células no serían exactamente iguales. Cuanto mayor fuese la proporción de mi cuerpo que se reemplazara, tanto menos como yo sería la persona resultante. En el caso situado en el extremo lejano de esta gama, todo mi cuerpo sería destruido, y harían una Réplica de una persona muy diferente, por ejemplo Greta Garbo. La Réplica de Garbo evidentemente *no* sería yo. En el caso situado en el extremo cercano, donde no habría sustitución alguna, la persona resultante sería yo. Según cualquier concepción, tiene que haber casos entre medias donde no podríamos responder a nuestra pregunta.

Por simplicidad, sólo voy a considerar el Espectro Físico, y voy a asumir que, en algunos de los casos de esta gama, no podemos responder a la pregunta de si la persona resultante sería yo. Mis observaciones podrían aplicarse, con algún ajuste, al Espectro Combinado.

Como he dicho, es natural asumir que, aunque *nosotros* no podamos responder a esta pregunta, siempre tiene que *existir* una respuesta, que tiene que ser Sí o No. Es natural pensar que si la persona resultante va a tener dolor, o yo sentiré ese dolor o no lo sentiré. Pero esta gama de casos desafía esa creencia. En el caso situado en el extremo cercano, la persona resultante sería yo. En el caso situado en el extremo lejano, sería alguien distinto. ¿Cómo podría ser cierto que, en todos los casos entre medias, esa persona

tiene que ser o yo o alguien distinto? Para que eso sea verdadero tiene que existir, en algún lugar de esta gama, una línea fronteriza nítida. Tiene que haber un conjunto crítico de células tal que, si sólo se las reemplaza a ellas, sería yo quien despertaría, pero tal que justo en el siguiente caso, con sólo unas cuantas más células reemplazadas, no sería yo sino una persona nueva. Eso es difícil de creer.

Aquí tenemos otro hecho que lo hace aún más difícil de creer. Aunque hubiera una línea fronteriza semejante, nadie podría descubrir jamás dónde se halla. Yo podría decir: «Trata de reemplazar la mitad de mi cerebro y de mi cuerpo y te diré lo que sucede». Pero sabemos por adelantado que, en todos los casos, como la persona resultante sería exactamente como yo, estaría inclinada a creer que era yo. Y esto no podría demostrar que ella *era* yo, puesto que toda simple Réplica mía pensaría eso también.

Aunque estos casos ocurrieran en realidad, no aprenderíamos nada más sobre ellos. De modo que no importa que sean imaginarios. Ahora deberíamos tratar de decidir si en esta gama de casos, la identidad personal podría ser determinada. ¿Podría ser verdadero que, en todos y cada uno de los casos, la persona resultante fuese yo o no fuese yo?

Si no creemos que haya Egos Cartesianos, u otras entidades semejantes, parece que estamos obligados a responder No. No es verdadero que nuestra identidad tenga que ser determinada. Siempre podemos preguntar: «¿Sería yo esa persona futura?». Pero, en algunos casos de estos,

(7) Esta pregunta no tendría respuesta. No sería ni verdadero ni falso que esta persona fuese yo.

Y

(8) Esta pregunta sería *vacía*. Aunque no tuviésemos una respuesta podríamos conocer toda la verdad sobre lo que ocurrió.

Si nuestras preguntas versaran sobre entidades tales como las naciones o las máquinas la mayoría de nosotros aceptaría estas afirmaciones. Pero cuando nos las aplicamos a nosotros mismos pueden ser difíciles de creer. ¿Cómo no iba a poder ser ni verdadero ni

falso que yo vaya a existir todavía mañana? Y sin una respuesta a nuestra pregunta ¿cómo iba a poder saber yo toda la verdad sobre mi futuro?

El Reduccionismo nos da la explicación. Asumimos de manera natural que en estos casos hay diferentes posibilidades. La persona resultante, asumimos, podría ser yo o podría ser alguien distinto que simplemente es como yo. Si la persona resultante va a sufrir un dolor, o yo sentiré ese dolor o no lo sentiré. Si realmente estas fueran posibilidades diferentes, sería necesario que una de ellas tuviera que ser la posibilidad que de hecho se realizase. ¿Cómo iba la realidad a poder dejar de elegir entre ellas? Pero para una concepción reduccionista,

> (9) Nuestra pregunta no versa sobre diferentes posibilidades. Hay sólo una única posibilidad o curso de sucesos. Nuestra pregunta versa meramente sobre diferentes descripciones posibles de este curso de sucesos.

Así es como nuestra pregunta no tiene respuesta. Todavía no hemos decidido cuál descripción aplicar. Y por esa razón, aun sin responder a esta pregunta, podríamos conocer toda la verdad sobre lo que ocurriría.

Supongamos que, tras considerar ejemplos tales, dejamos de creer que nuestra identidad tenga que ser determinada. Eso puede parecer que representa una diferencia pequeña. Puede parecer que se trata de un cambio de concepción sólo en lo que respecta a algunos casos imaginarios que nunca ocurrirán en realidad. Pero puede que no sea así. Podemos ser llevamos a revisar nuestras creencias sobre la naturaleza de la identidad personal, y eso sería un cambio de concepción en lo que se refiere a nuestra propia vida.

En casi todos los casos reales, las preguntas sobre la identidad personal tienen respuesta, de forma que la afirmación (7) no se aplica. Si no conocemos estas respuestas hay algo que no sabemos. Pero la afirmación (8) todavía se aplica. Aun sin responder a estas preguntas podríamos conocer toda la verdad de lo que ocurre. Sabríamos esa verdad si conociéramos los hechos acerca de la continuidad tanto física como psicológica. Si, lo que resulta poco vero-

símil, todavía no conociéramos la respuesta a una pregunta sobre la identidad, nuestra ignorancia sólo sería ignorancia acerca de nuestro lenguaje. Y eso es porque la afirmación (9) todavía se aplica. Cuando conocemos los otros hechos, nunca hay diferentes posibilidades en el nivel de lo que sucede. En todos los casos, las únicas posibilidades que quedan se hallan en el nivel lingüístico. Tal vez sería correcto decir que alguna persona futura sería yo. Tal vez sería correcto decir que no sería yo. O tal vez ninguna de las dos maneras de hablar sería correcta. Concluyo que en *todos* los casos, si conocemos los otros hechos, deberíamos considerar las preguntas sobre nuestra identidad como meras preguntas sobre el lenguaje.

Esta conclusión puede malentenderse. En primer lugar, cuando hacemos tales preguntas normalmente es porque *no* conocemos los otros hechos. Así, cuando preguntamos si estamos a punto de morir, rara vez se trata de una cuestión conceptual. Hacemos esa pregunta porque no sabemos lo que le va a ocurrir a nuestro cuerpo, ni si, sobre todo, nuestro cerebro va a seguir sirviendo de soporte a la conciencia. Nuestra pregunta se convierte en conceptual sólo cuando ya conocemos lo que se refiere a estos otros hechos.

Hay que tomar nota además de que, en ciertos casos, los hechos relevantes van más allá de los detalles del caso que estamos considerando. Que se aplique un concepto puede depender de hechos acerca de otros casos, o de una elección entre teorías científicas. Supongamos que vemos que algo raro le ocurre a un animal desconocido. Podríamos preguntar si este proceso preserva la identidad del animal, o si el resultado es un animal nuevo (porque lo que estamos viendo sea alguna clase de reproducción). Aunque conociéramos los detalles de este proceso, esa pregunta no sería meramente conceptual. La respuesta dependería de si este proceso es parte del desarrollo natural de esta especie animal. Y eso puede ser algo que todavía tengamos que descubrir.

Si identificamos a las personas con los seres humanos, a los que consideramos como una clase natural, lo mismo podría ocurrir en algunos casos imaginarios que implicasen a personas. Pero estos no son el tipo de caso que he estado discutiendo. Todos mis casos conllevan la intervención artificial. Aquí no podría ser rele-

vante ningún hecho sobre el desarrollo natural. Así, en mi Espectro Físico, si sabemos cuáles de mis células serían reemplazadas por duplicados, la totalidad de los hechos empíricos relevantes estaría incluida. En casos así, cualquier pregunta que quedara sería conceptual.

Siendo así las cosas, sería más claro hacer estas preguntas de un modo diferente. Consideremos el caso en que yo sustituyo algunos de mis componentes de mi equipo de audio pero conservo otros. Pregunto: «¿Tengo todavía el mismo equipo de audio?». Esta puede parecer una pregunta fáctica. Pero como yo ya sé lo que ocurrió, no es así en realidad. Sería más claro preguntar: «Dado que he reemplazado esos componentes, ¿sería correcto llamarlo el mismo equipo de audio?».

Lo mismo se aplica a la identidad personal. Supongamos que yo conozco los hechos sobre lo que le va a ocurrir a mi cuerpo, y sobre todas las conexiones psicológicas que habrá entre yo ahora y una persona mañana. Puedo preguntar: «¿Seré yo esa persona?». Pero esta es una manera de formular mi pregunta que induce confusión. Sugiere que no sé lo que va a ocurrir. Cuando conozco estos otros hechos, debería preguntar: «¿Sería correcto llamar yo a esa persona?». Eso me recordaría que, si hay algo que yo no conozco, se trata simplemente de un hecho sobre nuestro lenguaje.

Creo que podemos ir más lejos. Tales preguntas son, en sentido peyorativo, meramente verbales. Hay cuestiones conceptuales que sin duda merece la pena discutir. Pero las cuestiones sobre la identidad personal, en los tipos de caso que analizo, son como esas cuestiones que todos consideraríamos sin importancia. No tiene el más mínimo interés si todavía tengo el mismo equipo de audio con la mitad de sus componentes sustituidos. De la misma manera deberíamos considerar sin el más mínimo interés la pregunta de si yo todavía existiría en caso de que la mitad de mi cuerpo fuese reemplazada de una vez. Como preguntas acerca de la realidad, estas son completamente vacías. Ni tampoco necesitan respuesta en cuanto preguntas conceptuales.

Pero podríamos necesitar, para propósitos legales, *dar* respuesta a preguntas semejantes. Así, podríamos decidir que un equipo de

audio debería llamarse el mismo si sus nuevos componentes cuestan menos de la mitad de su precio original. Y podríamos decidir decir que yo seguiría existiendo en la medida en que menos de la mitad de mi cuerpo sea reemplazada. Pero estas no son respuestas a preguntas conceptuales; son meras decisiones.

(Observaciones similares son aplicables si somos reduccionistas identificadores, los que creen que las personas son sólo cuerpos. Hay casos en que es una cuestión meramente verbal la de si todavía tenemos el mismo cuerpo humano. Esto es sin duda cierto en los casos situados en la mitad del Espectro Físico.)

Puede ser útil contrastar estas preguntas con una que no es meramente verbal. Supongamos que estamos estudiando una criatura que es muy distinta de nosotros, por ejemplo un insecto o algún ser extraterrestre. Conocemos todos los hechos sobre la conducta de esta criatura, así como su neurofisiología. De repente la criatura se revuelve enérgicamente, en lo que parece ser una respuesta a un daño. Preguntamos: «¿Es consciente, y está sufriendo un gran dolor? ¿O es simplemente como una máquina sin conciencia?». Un conductista podría decir: «Esa es una pregunta meramente verbal. Estas no son posibilidades diferentes, de las que una podría ser la verdadera. Son simplemente descripciones diferentes del mismo estado de hechos». Eso yo lo encuentro increíble. Estas descripciones nos dan, creo, dos posibilidades muy diferentes. No podría ser una cuestión vacía o meramente verbal la de si una criatura es inconsciente o sufre un gran dolor.

Es natural pensar lo mismo de nuestra propia identidad. Si yo sé que cierta proporción de mis células va a ser reemplazada, ¿cómo puede ser una cuestión meramente verbal la de si estoy a punto de morir, o por el contrario me despertaré de nuevo mañana? Es porque eso es difícil de creer por lo que el Reduccionismo vale la pena discutirlo. Si nos volvemos reduccionistas pueden cambiar algunas de nuestras más profundas asunciones acerca de nosotros mismos.

Estas asunciones, como he dicho, cubren casos reales y nuestra propia vida. Pero son puestas de manifiesto de la mejor manera cuando consideramos casos problema imaginarios. Merece la pena explicar con más detalle el porqué.

En los casos corrientes las preguntas sobre nuestra identidad tienen respuesta. En esos casos hay un hecho referente a la identidad personal, y el Reduccionismo es una concepción acerca de qué clase de hecho es este. Según ella, la identidad personal no consiste más que en continuidad física y/o psicológica. Podemos encontrar difícil de decidir si aceptamos este modo de ver las cosas, puesto que puede ser muy dudoso cuándo un hecho no consiste sino en otro. Podemos hasta dudar de si los reduccionistas y sus críticos están realmente en desacuerdo.

En los casos problema las cosas son diferentes. Cuando no podemos responder a preguntas sobre la identidad personal, es más fácil decidir si aceptamos una concepción reduccionista. Deberíamos preguntar: ¿Encontramos desconcertantes estos casos? ¿O aceptamos la afirmación reduccionista de que, hasta sin contestar estas preguntas, si conocemos los hechos sobre las continuidades entonces conoceríamos lo que ocurrió?

La mayoría de nosotros encuentra estos casos desconcertantes. Creemos que, aunque conociéramos esos otros hechos, si no pudiéramos responder preguntas sobre nuestra identidad habría algo que no conoceríamos. Eso sugiere que, según nuestro modo de pensar, la identidad personal *no* consiste sólo en una o dos de esas continuidades, sino que es un hecho que se da separadamente, una diferencia adicional en lo que ocurre. La explicación reduccionista tiene entonces que dejar algo fuera. De modo que hay un desacuerdo real, uno que es aplicable a todos los casos.

Muchos de nosotros no nos limitamos a encontrar desconcertantes esos casos. Estamos inclinados a pensar que, en todos los casos como estos, las preguntas referentes a nuestra identidad tienen que tener respuesta, y una respuesta que tiene que ser Sí o No. Para que eso sea así, la identidad personal tiene que ser un hecho que se da separadamente, de una clase especialmente simple. Tiene que implicar a una entidad especial, tal como un Ego Cartesiano, cuya existencia tiene que ser todo-o-nada.

Cuando digo que tenemos estos supuestos, *no* estoy afirmando que creamos en Egos Cartesianos. Algunos de nosotros sí. Pero me temo que muchos de nosotros tenemos creencias inconsistentes. Si

se nos pregunta si creemos que hay Egos Cartesianos, tal vez respondamos No. Y puede que aceptemos que, como afirman los reduccionistas, la existencia de una persona conlleva sólo la existencia de un cuerpo y la ocurrencia de una serie de sucesos mentales y físicos interrelacionados. Pero como demuestran nuestras reacciones a los casos problema, nosotros no aceptamos del todo esta concepción. O, si lo hacemos, también parecemos sostener una concepción diferente.

Tal conflicto de creencias es muy común. A un nivel reflexivo o intelectual podemos hallarnos convencidos de que una concepción es verdadera; pero en otro nivel, ese que tiene que ver más directamente con nuestras emociones, podemos seguir pensando y sintiendo como si fuese verdadera una concepción diferente. Un ejemplo de este tipo sería una esperanza o un temor que sabemos que carece de fundamento. Me temo que muchos de nosotros tenemos creencias inconsistentes como estas cuando nos ponemos a reflexionar sobre los temas centrales de la Metafísica: Dios, el Yo, la conciencia, el tiempo y el libre albedrío.

II

Me vuelvo ahora desde la naturaleza de la identidad personal a su importancia. Mucha gente piensa que la identidad personal tiene una gran significación racional y moral. Así, se piensa que es el hecho de la identidad lo que nos da la razón que tenemos para preocuparnos por nuestro propio futuro. Y varios principios morales, como los del merecimiento o la justicia distributiva, presuponen afirmaciones sobre la identidad. Se ha denominado a la condición separada de las personas, o a la no identidad de diferentes personas, «el hecho básico de la moral».

Hoy sólo me da tiempo a comentar una de estas cuestiones: qué es lo que importa en nuestra supervivencia. Me refiero con ello no a qué hace nuestra supervivencia buena, sino a qué hace que nuestra supervivencia importe, tanto si es buena como si es mala. ¿Qué es aquello, en nuestra supervivencia, que nos da una razón para una preocupación anticipatoria o prudencial especial?

Podemos explicar esta pregunta con un caso imaginario extremo. Supongamos que, aunque me preocupo por mi futuro entero, estoy especialmente preocupado por lo que me vaya a ocurrir los martes futuros. Antes que sufrir un leve dolor un martes futuro yo preferiría un dolor severo cualquier otro día del futuro. El patrón de intereses sería irracional. El hecho de que un dolor sobrevenga un martes no es razón para que nos preocupe más. ¿Y qué sucede con el hecho de que el dolor vaya a ser *mío*? ¿Me proporciona *este* hecho una razón para preocuparme más por él?

Mucha gente respondería Sí. Según su modo de pensar, lo que nos da una razón para preocuparnos por nuestro futuro es precisamente que va a ser nuestro futuro. La identidad personal es lo que importa en la supervivencia.

Yo rechazo esta concepción. Lo que sobre todo importa, pienso, son otras dos relaciones: la continuidad y la conexividad psicológicas que, en los casos corrientes, se dan entre las diferentes partes de la vida de una persona. Estas relaciones sólo coinciden aproximadamente con la identidad personal, dado que, a diferencia de ella, son en parte cuestión de grado. Tampoco pienso que importen tanto como se piensa que importa la identidad.

Hay diferentes maneras de cuestionar la importancia de la identidad.

Un argumento puede resumirse así:

(1) La identidad personal no consiste más que en ciertos otros hechos.
(2) Si un hecho no consiste más que en ciertos otros, sólo pueden ser estos otros hechos los que tengan importancia racional o moral. Deberíamos preguntar si, en sí mismos, estos otros hechos importan.

Por consiguiente

(3) La identidad personal no puede ser importante ni racional ni moralmente. Lo que importa sólo puede ser uno o más de los otros hechos en los que consiste la identidad personal.

La premisa (1) es Reduccionismo; a la (2) la podríamos llamar «Reduccionismo de la Importancia».

Mark Johnston critica este argumento [4]. Lo califica como un *Argumento desde Abajo*, puesto que el mismo afirma que, si un hecho consiste sólo en ciertos otros hechos, sólo pueden ser estos otros hechos de nivel más bajo los que importen. Johnston replica con lo que denomina un *Argumento desde Arriba*. Según su modo de ver la cuestión, aunque los hechos del nivel inferior no importen en sí mismos, el hecho de nivel superior sí que puede importar. Si importa, los hechos de nivel inferior tendrán una significación derivada. Ellos importarán, no en sí mismos, sino porque constituyen el hecho de nivel superior.

Para ilustrar este desacuerdo podemos comenzar con un caso diferente. Supongamos que preguntamos qué queremos que ocurra si, a consecuencia de una lesión cerebral, quedamos sin conciencia de una manera irreversible. Si estuviéramos en este estado, todavía estaríamos con vida. Pero este hecho debería entenderse de una manera reduccionista. No puede ser lo mismo que el hecho de que nuestro corazón todavía latiría y nuestros otros órganos aún estarían funcionando. Pero no sería un hecho independiente o que se diera de forma separada. Nuestro estar todavía con vida, aunque irreversiblemente inconscientes, no consistiría sino en estos otros hechos.

Según mi Argumento desde Abajo, deberíamos preguntarnos si, en sí mismos, esos otros hechos importan. Si nos hubiéramos quedado sin conciencia de un modo irreversible, ¿sería bueno para nosotros o para los demás que nuestro corazón y otros órganos todavía estuvieran en funcionamiento? Si respondemos No, deberíamos concluir que no importaría que todavía estuviéramos vivos.

Si Johnston tuviera razón podríamos rechazar este argumento y apelar a un Argumento desde Arriba. Podríamos decir:

[4] En su «Reasons and Reductionism» [«Razones y Reduccionismo»], en *Derek Parfit and His Critics*, editado por Jonathan Dancy, Blackwell, de próxima publicación. (Apareció en 1997. N. del t.)

> Puede que no sea en sí mismo bueno que nuestro corazón y otros órganos estuvieran todavía funcionando. Pero es bueno estar vivo. Siendo así las cosas, es racional esperar que, aunque nunca pudiésemos recuperar la conciencia, nuestro corazón siga latiendo tanto tiempo como sea posible. Eso sería bueno porque constituiría nuestro seguir con vida.

Creo que, de estos argumentos, el mío es el más admisible.

Consideremos a renglón seguido la cuestión moral que estos casos hacen surgir. Hay personas que piden, en su Última Voluntad, que si una lesión cerebral las dejara irreversiblemente inconscientes, su corazón sea detenido. Yo creo que debemos hacer lo que piden estas personas. Pero muchos adoptan una posición diferente. Podrían apelar a un Argumento desde Arriba. Podrían decir:

> Aunque tales personas nunca puedan recuperar la conciencia, puede decirse de ellas con verdad que siguen vivas mientras su corazón todavía late. Siendo así las cosas, detenerles el corazón sería un acto de matar. Y a no ser en defensa propia, matar a alguien es siempre incorrecto.

Según esta manera de pensar, debemos dejar que el corazón de estas personas siga latiendo durante meses o incluso años.

Como respuesta a la cuestión moral, esta me parece equivocada. (Es una cuestión diferente cuál debería ser la ley.) Pero, para muchas personas, la palabra «matar» tiene tal fuerza que parece significativo el que sea aplicable.

Volvámonos ahora a un tema diferente. Supongamos que, después de tratar de decidir cuándo tienen las personas libre albedrío, hemos quedado convencidos por cualquiera de dos concepciones compatibilistas. Según una de ellas, llamamos «no libres» a las elecciones si están causadas de ciertos modos, y las llamamos «libres» si están causadas de otros ciertos modos. Según la otra concepción, llamamos «no libres» a las elecciones si sabemos cómo fueron causadas, y las llamamos «libres» si todavía no lo hemos descubierto.

Supongamos a renglón seguido que, una vez que consideramos estos dos fundamentos para trazar esta distinción, llegamos a pensar que ninguno, en sí mismo, tiene la clase de significación que

podría dar apoyo a hacer o a negar declaraciones sobre la culpabilidad o el merecimiento. Nos parece que no hay tal significación en la diferencia entre estas clases de determinación causal, y somos de la opinión de que no puede importar que las causas de una decisión hayan sido ya descubiertas. (Nótese que, al comparar los Argumentos desde Arriba y desde Abajo, no tenemos necesidad de aceptar realmente estas afirmaciones. Estamos preguntando si, *en caso de que* aceptemos las premisas relevantes, debemos ser persuadidos por estos argumentos.)

Según mi Argumento desde Abajo, si el hecho de que una elección es libre no consiste sino en uno de esos otros hechos, y creemos que esos otros hechos no pueden ser moralmente importantes en sí mismos, deberíamos concluir que no puede ser importante el que la elección de una persona fuese libre. O las elecciones que no son libres pueden merecer castigo, o las elecciones que son libres no pueden merecerlo. Según un Argumento desde Arriba de tipo johnstoniano, aunque esos otros hechos no sean en sí mismos importantes —aunque en sí mismos sean triviales— pueden tener una importancia derivada si y porque constituyen el hecho de que la elección de una persona fue libre. Como antes, el Argumento desde Abajo me parece más admisible.

Ahora podemos considerar la cuestión subyacente en torno a la que gira este desacuerdo.

Como he afirmado, si un hecho no consiste sino en otros, el primer hecho no es un hecho independiente o que se dé de forma separada. Y, en los casos que ahora nos conciernen, es además, en relación con estos otros hechos, simplemente un hecho conceptual. Así, si alguien está inconsciente de manera irreversible pero su corazón sigue latiendo, es un hecho conceptual el que esta persona todavía está viva. Cuando le llamo conceptual a este hecho, no quiero decir que sea un hecho sobre nuestros conceptos. Que esta persona está viva es un hecho sobre esta persona. Pero si ya hemos dicho que su corazón todavía late, cuando afirmamos que la persona todavía está viva no damos información suplementaria sobre la realidad. Sólo damos información suplementaria sobre nuestro uso de las palabras «persona» y «viva».

Cuando volvemos a preguntar qué importa, la cuestión central es esta. Supongamos que estamos de acuerdo en que no importa en sí mismo que el corazón de una tal persona todavía esté latiendo. ¿Podríamos sostener que, de otra manera, este hecho sí que importa, porque hace correcto decir que esta persona se halla todavía con vida? Si contestamos Sí estamos tratando al lenguaje como más importante que la realidad. Estamos afirmando que, aunque un hecho no importe en sí mismo, puede importar si y porque permite que sea aplicada una determinada palabra.

Creo que esto es irracional. Según mi parecer, lo que importa son los hechos del mundo, los cuales, una vez dados, hacen que un concepto tenga aplicación. Si los hechos del mundo no tienen significación racional ni moral, y el hecho al que el concepto se aplica no supone una diferencia más en lo que ocurre, este hecho conceptual no puede ser significativo.

Johnston formula una segunda acusación contra el Reduccionismo de la Importancia. Si el Fisicalismo fuese verdadero, afirma él, todos los hechos no consistirían sino en hechos acerca de las partículas fundamentales. Considerados en sí mismos, estos hechos acerca de las partículas no tienen importancia racional ni moral. Si aplicamos el Reduccionismo de la Importancia tenemos que concluir que nada tiene ninguna importancia. Y este autor comenta: «esto no es una demostración del Nihilismo, es una *reductio ad absurdum*».

Teniendo en cuenta lo que he dicho hoy, esta acusación, creo, puede ser contestada. Puede que haya un sentido en que, si el Fisicalismo fuese verdadero, todos los hechos no consistirían sino en hechos acerca de las partículas fundamentales. Pero no es esa la clase de reducción que yo tenía en mente. Cuando afirmo que la identidad personal no consiste más que en ciertos otros hechos, tengo en mente una relación más íntima, y en parte conceptual. Las afirmaciones sobre la identidad personal pueden no significar lo mismo que las declaraciones sobre la continuidad física y/o psicológica. Pero si conociéramos los hechos acerca de estas continuidades, y comprendiésemos el concepto de persona, con ello conoceríamos, o podríamos llegar a entender, los hechos acerca de las

personas. De ahí mi afirmación de que, si conocemos los otros hechos, las preguntas sobre la identidad personal deberían ser tomadas como preguntas no sobre la realidad sino sobre nuestro lenguaje. Estas afirmaciones no son aplicables a los hechos acerca de las partículas elementales. Por ejemplo, no es verdad que si supiésemos cómo se mueven las partículas en el cuerpo de una persona, y comprendiéramos nuestros conceptos, con ello conoceríamos, o podríamos llegar a entender, todos los hechos relevantes acerca de esta persona. Para comprender el mundo que nos rodea nos hace falta más que la Física y un conocimiento de nuestro propio lenguaje. Necesitamos Química, Biología, Neurofisiología, Psicología, y muchas más cosas, además.

Si somos reduccionistas de la importancia, lo que no tenemos por qué afirmar es que, cuandoquiera que haya hechos en diferentes niveles, sea siempre el nivel más bajo el que importe. Esto es claramente falso. Estamos discutiendo casos en que, en relación a los hechos de un nivel inferior, el hecho de nivel superior es, en el sentido que he bosquejado, meramente conceptual. Nuestra tesis es que tales hechos conceptuales no pueden ser racional ni moralmente importantes. Lo que importa es la realidad, no cómo es descrita. (De modo que esta concepción podría llamarse mejor *Realismo* de la Importancia.)

He descrito ahora brevemente el Reduccionismo de las personas y el Reduccionismo de la Importancia. Los dos juntos implican que la identidad personal no es lo que importa. ¿Debemos aceptar esta conclusión?

La mayor parte de nosotros piensa que debemos preocuparnos de nuestro futuro porque va a ser *nuestro* futuro. Yo pienso que lo que importa no es la identidad sino otras determinadas relaciones. Para ayudarnos a decidir entre estas concepciones deberíamos considerar casos en que la identidad y esas relaciones no coinciden. Cuáles son estos casos depende de cuál criterio de identidad aceptamos. Comenzaré con la forma más simple del Criterio Físico, de acuerdo con la cual una persona sigue existiendo si y sólo si el cuerpo de esa persona sigue existiendo. Esa tiene que ser la opinión de los que creen que las personas no son sino cuerpos. Y es la opinión

de algunos de los que identifican a las personas con los seres humanos [5]. Llamémosla el *Criterio Corporal.*

Discuto esta concepción por una razón especial. Como veremos, hay otro argumento a favor de la falta de importancia de la identidad, el que apela al caso imaginario de la División, de Wiggins. Pero aquellos que aceptan el Criterio Corporal rechazan una premisa de ese otro argumento. Para persuadirlos de que la identidad no es lo que importa, mi único argumento es el Reduccionismo de la Importancia.

Supongamos que, a consecuencia de una lesión espinal, me he quedado parcialmente paralítico. Tengo un hermano que se muere de una enfermedad cerebral. Con ayuda de nuevas técnicas, cuando el cerebro de mi hermano deje de funcionar, mi cabeza podría injertarse en el resto de su cuerpo. Como somos gemelos idénticos, mi cerebro, en ese caso, controlaría un cuerpo que es exactamente como el mío, salvo que no estaría paralítico.

¿Debería aceptar esta operación? De los que asumen que la identidad es lo que importa, tres grupos contestarían No. Unos aceptan el Criterio Corporal. Estos piensan que si se realizara la operación yo moriría. La persona con mi cabeza mañana sería mi hermano, que pensaría equivocadamente que él era yo. Otros no están seguros de lo que sucedería. Creen que sería arriesgado aceptar la operación, puesto que la persona resultante podría no ser yo. Otros dan una razón diferente de por qué yo debería rechazar esta operación: que sería indeterminado el que esa persona fuese yo. Según todas estas maneras de ver el asunto, lo que importa es quién sería esa persona.

Según mi parecer, la cuestión carece de importancia. Si se llevase a cabo esta operación, la persona con mi cabeza mañana no sólo creería que era yo, parecería recordar haber vivido mi vida y sería en todo otro respecto como yo desde el punto de vista psicológico. Estos hechos también tendrían su causa normal, la continua existencia de mi cerebro. Y el cuerpo de esta persona sería exactamente como el mío. Por todas estas razones, su vida sería exactamente como la vida que yo habría vivido si se hubiera curado mi parálisis. Creo que, dados estos hechos, yo debería aceptar la operación. Es irrelevante que esta persona vaya a ser yo o no.

[5] Véase, por ejemplo, Michael Ayers, «Locke», vol. II.

Eso puede parecer importantísimo. Después de todo, si ella no fuese a ser yo, yo habré dejado de existir. Pero si esa persona no fuese a ser yo, este hecho no consistiría más que en otro hecho. No consistiría sino en el hecho de que mi cuerpo habrá sido reemplazado del cuello para abajo. Una vez considerado en sí mismo, ¿es importante ese segundo hecho? ¿Puede importar en sí mismo que la sangre que va a mantener vivo mi cerebro no circulará por mi propio corazón y mis propios pulmones, sino por el corazón y los pulmones de mi hermano? ¿Puede importar de suyo que mi cerebro vaya a controlar no el resto de mi cuerpo sino el resto de otro cuerpo que es exactamente igual?

Si creemos que estos hechos equivaldrían a mi inexistencia, puede ser difícil centrarse en la cuestión de si, en ellos mismos, estos hechos importan. Para hacerlo más fácil, deberíamos imaginar que aceptamos una concepción diferente. Supongamos que estamos convencidos de que la persona con mi cabeza mañana *sería* yo. ¿Creeríamos entonces que importaría mucho que mi cabeza hubiese sido injertada en este otro cuerpo? No lo creeríamos. Consideraríamos el hecho de recibir por mi parte un nuevo torso y nuevos miembros como algo semejante a un transplante menor, algo parecido a recibir un corazón nuevo, o unos nuevos riñones. Como esto muestra, si importara mucho que lo que va a ser reemplazado no es sólo unos pocos órganos sino todo mi cuerpo del cuello para abajo, eso sólo podría ser porque, si eso ocurriera, la persona resultante *no* sería yo.

De acuerdo con el Reduccionismo de la Importancia, deberíamos concluir ahora que ninguno de estos hechos podría importar mucho. Como no sería de suyo importante que mi cabeza fuera injertada en este cuerpo, y eso sería todo lo que hubiese en el hecho de que la persona resultante no fuese a ser yo, no sería en sí mismo importante que esta persona no fuese a ser yo. Tal vez no sería irracional lamentar estos hechos un poco. Pero creo que serían superados con mucho por el hecho de que, a diferencia de mí, la persona resultante no sería paralítica.

Cuando se aplica a nuestra propia existencia, el Reduccionismo de la Importancia es difícil de creer. Pero, igual que antes, la cuestión fundamental es la importancia relativa del lenguaje y la realidad.

Según mi concepción lo que importa es lo que va a suceder. Si yo supiera que mi cabeza podría injertarse en el resto de un cuerpo que es exactamente como el mío, y que la persona resultante sería exactamente igual que yo, sabría lo bastante como para decidir si acepto esta operación. No necesito preguntar si la persona resultante podría ser llamada correctamente yo. Eso no es una diferencia ulterior en lo que va a suceder.

Eso puede parecer una falsa distinción. Lo que importa, podríamos decir, es si la persona resultante *sería* yo. Pero esa persona sería yo si y sólo si pudiera ser llamada correctamente yo. De manera que, al preguntar cómo podría ser llamada, no estamos haciendo meramente una pregunta conceptual. *Preguntamos* sobre la realidad.

Esta objeción falla al no distinguir dos tipos de casos. Supongamos que le pregunto a mi médico si me dolerá recibir un tratamiento. Esa es una pregunta fáctica. Pregunto qué va a suceder. Como el dolor puede ser llamado «dolor», yo *podría* hacer mi pregunta de otra manera. Podría decir: «Mientras esté bajo tratamiento, ¿será correcto describirme como sufriendo dolor?». Pero eso induciría a confusión. Daría a entender que estoy preguntando cómo usamos la palabra «dolor».

En un caso diferente yo podría hacer esa pregunta conceptual. Supongamos que sé que cuando cruce el Canal me voy a marear en el barco porque siempre me pasa. Podría preguntarme si se podría llamar correctamente «dolor» a esa sensación. Aquí también podría hacer la pregunta de otra manera. Podría decir, «Al cruzar el canal, ¿me dolerá?». Pero eso induciría a confusión puesto que daría a entender que pregunto por lo que va a ocurrir.

En el caso médico, yo no sé en qué estado consciente voy a estar. Hay diferentes posibilidades. En el caso del paso del Canal, no hay posibilidades diferentes. Yo ya sé en qué estado estaré. Simplemente pregunto si ese estado podría redescribirse de un cierto modo.

Sí que tiene importancia que al recibir el tratamiento médico me vaya a doler o no. Y sí que tiene importancia que al cruzar el

canal me vaya a marear o no. Pero no la tiene que al marearme en el barco se pueda decir o no que me está doliendo.

Ahora volvamos a nuestro ejemplo principal. Supongamos que sé que mi cabeza va a ser injertada con éxito en el cuerpo decapitado de mi hermano. Yo pregunto si la persona resultante será yo. ¿Este es como el caso médico, o como el de cruzar el Canal? ¿Estoy preguntando qué ocurrirá, o si lo que sé que ocurrirá podría describirse de una cierta manera?

Según mi parecer, debería entender que estoy preguntando lo segundo. Yo ya sé lo que va a ocurrir. Habrá alguien con mi cabeza y con el cuerpo de mi hermano. Es una mera cuestión verbal la de si esa persona será yo. Y por esa razón, aunque no vaya a ser yo, eso no importa.

Puede objetarse ahora: «Al elegir este ejemplo estás haciendo trampas. Por supuesto que deberías aceptar la operación. Pero eso es porque la persona resultante *sería* tú. Deberíamos rechazar el Criterio Corporal. De modo que este caso no puede demostrar que la identidad no es lo que importa».

No hago trampas, porque hay gente que acepta este criterio. Vale la pena tratar de mostrar a estas personas que la identidad no es lo que importa. Pero acepto parte de esta objeción. Admito que deberíamos rechazar el Criterio Corporal. De los que apelan a este criterio, hay quienes creen que las personas no son sino cuerpos. Pero, si mantenemos esta clase de concepción sería mejor identificar a la persona con su cerebro o su sistema nervioso [6]. Consideremos a renglón seguido a los que creen que las personas son animales de un cierto tipo, *v. gr.* seres humanos. Nosotros podríamos adoptar este modo de pensar pero rechazar el Criterio Corporal. Podríamos afirmar que los animales siguen existiendo si siguen existiendo y funcionando las partes más importantes de sus cuerpos. Y podríamos afirmar que, al menos en el caso de los seres humanos, el cerebro es tan importante que su supervivencia equiva-

[6] Véase, por ejemplo, Thomas Nagel, *The View from Nowhere* [*La visión de ningún lugar*], OUP, 1986, pp. 40-45, y John Mackie, *Problems from Locke* [*Problemas a partir de Locke*], OUP, 1976, capítulo 6.

le a la supervivencia de este ser humano. Según estas dos concepciones, en mi caso imaginario la persona con mi cabeza mañana sería yo. Y eso es lo que, después de reflexionar, la mayor parte de nosotros creería.

Mi propia concepción es parecida. Yo la formularía no como una afirmación sobre la realidad sino como una de tipo conceptual. Según ella, no sería incorrecto llamar yo a esta persona; y esta sería la mejor descripción de este caso.

Si estamos de acuerdo en que esta persona sería yo, yo todavía sostendría que este hecho no es lo que importa. Lo que es importante no es la identidad, sino uno o más de los otros hechos en que la identidad consiste. Pero admito que cuando la identidad coincide con estos otros hechos es más difícil decidir si aceptamos la conclusión del argumento. De modo que si rechazamos el Criterio Corporal tenemos que considerar otros casos.

Supongamos que aceptamos la versión basada en el cerebro del Criterio Psicológico. Según esta concepción, si va a haber una persona futura que es psicológicamente continua conmigo, porque tendrá lo bastante de mi cerebro, esa persona será yo. Pero la continuidad psicológica sin su causa normal, la existencia continuada de suficiente cerebro mío, no basta para la identidad. Mi Réplica no sería yo.

Recordemos además que un objeto puede seguir existiendo aunque todos sus componentes sean sustituidos gradualmente. Supongamos que cada vez que un barco de madera entra en el puerto se reemplazan unas cuantas de sus tablas. Antes de que pase mucho tiempo ese mismo barco puede estar compuesto enteramente de tablas diferentes.

Supongamos una vez más que yo necesito cirugía. Todas las células de mi cerebro tienen un defecto que, andando el tiempo, sería fatal. Los cirujanos podrían reemplazar todas estas células, insertando células nuevas que son réplicas exactas, salvo que no tienen el defecto.

Los cirujanos podrían proceder de dos maneras. En el *Caso Uno* habría cien operaciones. En cada operación, los cirujanos quitarían una centésima parte de mi cerebro, e insertarían réplicas de esas partes, En el *Caso Dos*, los cirujanos quitarían primero todas las partes existentes de mi cerebro y luego insertarían todas sus réplicas.

Hay aquí una diferencia real. En el Caso Uno, mi cerebro seguiría existiendo, como un barco con todas sus tablas gradualmente reemplazadas. En el Caso Dos, mi cerebro dejaría de existir, y a mi cuerpo se le daría un cerebro nuevo.

Esta diferencia, sin embargo, es mucho más pequeña que la que hay entre la supervivencia corriente y el teletransporte. En ambos casos, habrá más tarde una persona cuyo cerebro será exactamente como mi cerebro actual, pero sin los defectos, una persona que por consiguiente será psicológicamente continua conmigo. Y en *ambos* casos el cerebro de esta persona estará hecho de las mismísimas células nuevas, cada una de las cuales es una réplica de una de mis células existentes. La diferencia entre los casos es meramente el modo en que se insertan estas nuevas células. En el Caso Uno, los cirujanos alternan entre quitar e insertar. En el Caso Dos, hacen toda la eliminación antes de toda la inserción.

Para el Criterio basado en el cerebro, esta es la diferencia entre la vida y la muerte. En el Caso Uno, la persona resultante sería yo. En el Caso Dos *no* sería yo, de forma que yo dejaría de existir.

¿Puede importar esta diferencia? Volvamos a aplicar el Argumento desde Abajo. Esta diferencia consiste en el hecho de que, en vez de alternar entre eliminaciones e inserciones, el cirujano hace toda la eliminación antes de toda la inserción. Considerado en sí mismo, ¿esto puede importar? Creo que no. No pensaríamos que importa si no constituyera el hecho de que la persona resultante no sería yo. Pero si este hecho no importa de suyo, y eso es todo lo que hay en el hecho de que en el Caso Dos yo dejaría de existir, debería concluir que mi dejar de existir no importa.

Supongamos además que consideras estos casos como casos problema, casos en que no sabes qué me ocurriría. Volvamos al Espectro Físico, que es más simple. En cada uno de los casos en esta gama, alguna proporción de mis células será reemplazada por duplicados exactos. Con algunas proporciones —20%, digamos, o 50%, o 70%— la mayoría de nosotros no estaríamos seguros de si la persona resultante sería yo. (Como antes, si no creo eso aquí, mis

comentarios podrían transferirse, con algunos ajustes, al Espectro Combinado.)

Según la concepción que mantengo, en todos los casos de esta gama es una cuestión meramente conceptual la de si la persona resultante sería yo. Aún sin responder a ella, yo puedo saber qué va a ocurrir exactamente. Si hay algo que no sé, se trata simplemente de un hecho acerca de cómo podríamos describir lo que va a ocurrir. Y considero que esa cuestión conceptual ni siquiera es interesante. Es meramente verbal, como la de si, en caso de que sustituyera algunas de sus partes, yo todavía tendría el mismo equipo de audio.

Cuando nos imaginamos estos casos desde el punto de vista de la primera persona, todavía puede ser difícil creer que esta sea meramente una cuestión verbal. Si no sé si aún existiré mañana, puede ser difícil de creer que sé lo que va a suceder. Pero, ¿qué es lo que no sé? Si hay diferentes posibilidades en el nivel de lo que sucede, ¿cuál es la diferencia entre ellas? ¿En qué consistiría esa diferencia? Si yo tuviera un alma, o Ego Cartesiano, podría haber diferentes posibilidades. Incluso si un *n* por ciento de mis células fuese reemplazado, tal vez mi alma mantendría su íntima relación con mi cerebro. O quizás tomaría el control otra alma. Pero hemos dado por descontado que no hay entidades semejantes. ¿Qué otra podría ser la diferencia? Cuando la persona resultante se despierte mañana, ¿qué podría hacer verdadero o falso que ella sea yo?

Puede decirse que al preguntar qué ocurrirá estoy preguntando qué puedo esperar. ¿Puedo esperar despertar de nuevo? Si esa persona sufrirá un dolor, ¿puedo esperar sentir ese dolor? Pero esto no sirve de ayuda. Estas no son más que otras maneras de preguntar si esa persona va a ser yo o no va a ser yo. Al apelar a lo que puedo esperar no explicamos qué produciría estas posibilidades diferentes.

Podemos pensar que esta diferencia no necesita explicación. Puede parecer que es suficiente decir: Quizás esa persona será yo y quizás no. Quizás yo existiré mañana, y quizás no. Puede parecer que estas tengan que ser posibilidades diferentes.

Sin embargo, eso es una ilusión. Si yo todavía voy a existir mañana, ese hecho tiene que consistir en ciertos otros. Para que haya dos posibilidades, de manera que pudiera ser verdadero o falso que voy a existir mañana, tiene que haber alguna otra diferencia entre estas posibilidades. Habría tal diferencia, por ejemplo, si, entre ahora y mañana, mi cerebro y mi cuerpo pudieran o permanecer indemnes o volar en pedazos. Pero en nuestro caso imaginario no hay tal otra diferencia. Ya sé que habrá alguien cuyo cerebro y cuyo cuerpo consistirán parcialmente en estas células y parcialmente en células nuevas, y que esta persona será psicológicamente como yo. No hay, en el nivel de lo que ocurre, diferentes resultados posibles. No hay una esencia adicional de mi yo o una propiedad de «yoidad», que podría o no podría estar allí.

Si nos volvemos al nivel conceptual, *hay* diferentes posibilidades. Quizás la persona futura podría ser correctamente llamada yo. Quizás podría ser correctamente llamada con el nombre de alguien distinto. O quizás ninguna de las dos formas sea correcta. Ese, sin embargo, es el único modo en que podría ser verdadero o falso que esta persona sería yo.

La ilusión puede persistir. Aun cuando yo conozca los otros hechos, puede que quiera que la realidad marche de una de dos formas. Puede que quiera que sea verdadero que yo vaya a existir mañana. Pero todo lo que podría ser verdadero es que usamos el lenguaje de una de dos formas. ¿Puede ser racional preocuparse por eso?

III

Ahora doy por sentado que aceptamos el Criterio Psicológico Basado en el Cerebro. Creemos que, si va a haber una persona futura que tendrá el suficiente cerebro mío para ser psicológicamente continua conmigo, esa persona sería yo. Según esta concepción, hay otro modo de argumentar que la identidad no es lo que importa.

En primer lugar, podemos tomar nota de que, igual que yo podría sobrevivir con menos que todo mi cuerpo, podría sobrevivir con menos que todo mi cerebro. Hay personas que han sobrevivido, y con poca alteración psicológica, aunque hayan perdido, por un ataque o una lesión, el uso de la mitad de su cerebro.

Supongamos además que cada una de las dos mitades de mi cerebro pudiera servir de soporte plenamente al funcionamiento psicológico corriente. Puede que eso ocurra de hecho con ciertas personas. Si no, podemos suponer que se ha hecho que ocurra conmigo merced a cierto avance tecnológico. Como nuestro objetivo es poner a prueba nuestras creencias sobre lo que importa, no hay ningún perjuicio en hacer tales asunciones.

Ahora podemos comparar dos operaciones posibles más. En la primera, una vez que se ha destruido la mitad de mi cerebro, la otra mitad sería transplantada con éxito en el interior de un cráneo vacío de un cuerpo que es exactamente como el mío. Dadas nuestras asunciones, deberíamos concluir que también aquí sobreviviría. Como yo sobreviviría si mi cerebro fuera transplantado y sobreviviría con sólo la mitad de mi cerebro, no sería razonable negar que yo sobreviviría si esa parte restante fuese transplantada. De manera que, en este *Caso de Una Cara*, la persona resultante sería yo.

Consideremos a continuación el *Caso de Dos Caras*, o *Mi División*. Las dos mitades de mi cerebro serían transplantadas con éxito en diferentes cuerpos que son exactamente como el mío. Se despertarían dos personas, cada una de ellas con la mitad de mi cerebro y cada una de ellas exactamente como yo, tanto física como psicológicamente.

Como estas serían dos personas diferentes, no puede ser cierto que cada una de ellas sea yo. Eso sería una contradicción. Si cada una de ellas fuera yo, cada una sería la misma persona: yo. De modo que no podrían ser dos personas diferentes.

¿Podría ser verdadero que sólo una de ellas sea yo? Eso no es una contradicción. Pero, como yo tengo la misma relación con cada una de estas personas, no hay nada que pudiera hacerme una de ellas en vez de la otra. No puede ser verdadero, de una u otra de estas personas, que sea ella la única que podría ser correctamente llamada yo.

¿Cómo debería considerar estas dos operaciones. ¿Preservarían lo que importa en la supervivencia? En el Caso de Una Cara, la única persona resultante sería yo. La relación entre yo ahora y esa persona futura es sólo un ejemplo de la relación entre yo ahora y yo mismo mañana. Así que esa relación contendría lo que importa. En el Caso de Dos Caras mi relación con esa persona sería justo la misma. De modo que esta tiene que contener todavía lo que importa. Nada se pierde. Pero no se puede afirmar aquí que esa persona sea yo. Así que la identidad no puede ser lo que importa.

Podemos objetar que si esa persona *no es* yo, algo *está* perdido. *Yo estoy* perdido. Puede parecer que eso supone toda la diferencia del mundo. ¿Cómo puede estar todo todavía ahí si *yo* no *estoy* ahí?

Todo está todavía ahí. El hecho de que yo no estoy ahí no es una ausencia real. La relación entre yo ahora y la persona futura es de suyo la misma. Como en el Caso de Una Cara, tiene la mitad de mi cerebro y es exactamente como yo. La diferencia es sólo que, en este Caso de Dos Caras, yo también tengo la misma relación con la otra persona resultante. ¿Por qué yo no estoy ahí? La explicación es sólo esta. Cuando esta relación se da entre yo ahora y una única persona en el futuro, se nos puede llamar la misma persona. Cuando esta relación se da entre yo ahora y *dos* personas futuras, a mí no me pueden llamar la misma persona que cada una de estas personas. Pero no es esa una diferencia en la naturaleza o el contenido de esta relación. En el Caso de Una Cara, donde la mitad de mi cerebro será transplantada con éxito, mi perspectiva es la supervivencia. Esa perspectiva contiene lo que importa. En el Caso de Dos Caras, donde las dos mitades serán transplantadas con éxito, nada se perdería.

Puede ser difícil de creer que la identidad no sea lo que importa. Pero es más fácil de aceptar cuando vemos por qué, en este ejemplo, es verdadero. Puede ser útil considerar esta analogía. Imaginemos una comunidad de personas que son como nosotros, pero con dos excepciones. Primera, como consecuencia de ciertos hechos referentes a su sistema reproductor, cada pareja tiene sólo dos hijos, que siempre son gemelos. Segundo, como consecuencia de ciertos rasgos especiales de su psicología, es de gran importancia para el desarrollo de cada niño que no se convierta en un hijo único como resultado de la muerte de su hermano. Tales niños sufren daño psicológico. Por eso se cree en esta comunidad que tiene mucha importancia que cada niño tenga un gemelo.

Supongamos ahora que, a causa de un cambio biológico, algunos de los niños de esta comunidad empiezan a nacer como trillizos. ¿Pensarían sus padres que esto es un desastre porque estos niños no tienen gemelos? Evidentemente no. Estos niños no tienen gemelos sólo porque cada uno de ellos tiene *dos* hermanos. Como cada niño tiene dos hermanos, el trío tiene que ser llamado trillizos y no gemelos. Pero ninguno de ellos sufrirá daño alguno como hijo único. Estas personas deberían revisar su modo de pensar. Lo que importa no es tener un gemelo, es tener al menos un hermano.

De la misma manera, nosotros deberíamos revisar nuestra concepción de la identidad a través del tiempo. Lo que importa no es que vaya a haber alguien vivo que será yo. Es más bien que vaya a haber al menos una persona viva que será psicológicamente continua conmigo como soy ahora, y/o que tiene suficiente cerebro mío. Cuando vaya a haber sólo una persona así, puede ser descrita como yo. Cuando vaya a haber dos personas así, no podemos afirmar que cada una será yo. Pero eso es tan trivial como el hecho de que si yo tuviera dos hermanos idénticos no podrían ser llamados mis gemelos.

IV

Si, como he argumentado, la identidad personal no es lo que importa, tenemos que preguntarnos qué importa. Hay varias respuestas posibles. Y, dependiendo de nuestra respuesta, hay varias implicaciones más. Así, hay varias cuestiones morales que no tengo tiempo siquiera de mencionar. Terminaré con otra observación acerca de nuestra preocupación por nuestro propio futuro.

Esa preocupación es de varias clases. Podemos querer sobrevivir, en parte, para realizar nuestras esperanzas y nuestras ambiciones. También podemos preocuparnos de nuestro propio futuro de la clase de forma en que nos preocupamos por el bienestar de ciertas otras personas, como nuestros parientes o nuestros amigos. Pero la mayoría de nosotros tiene, además, una clase distintiva de preocupación egoísta. Si yo sé que mi hijo va a sufrir dolor, puedo preocuparme por su dolor más de lo que me preocuparía por mi propio

dolor futuro. Pero no puedo anticipar, poseído por el miedo, el dolor de mi hijo. Y si yo supiera que mi Réplica reanudaría mi vida donde yo la dejo, no esperaría con ansia esa vida.

Esta clase de preocupación puede debilitarse, creo yo, y reconocerse sin fundamento, si llegamos a aceptar una Concepción Reduccionista. En nuestros pensamientos sobre nuestra propia identidad somos propensos a las ilusiones. Esa es la razón por la cual los denominados «casos problema» parecen levantar problemas: porque encontramos difícil de creer que, cuando conocemos los otros hechos, sea una cuestión vacía o meramente verbal la de si todavía existiremos. Aun tras aceptar una Concepción Reduccionista, podemos seguir pensando y sintiendo, en algún nivel, como si esa concepción no fuese verdadera. Nuestra propia existencia continua puede parecer aún un hecho independiente, de una clase particularmente profunda y simple. Y esa creencia puede subyacer a nuestra preocupación anticipadora por nuestro propio futuro.

Hay, sospecho yo, varias causas de esa creencia ilusoria. Hoy he discutido una de ellas: nuestro esquema conceptual. Aunque tenemos necesidad de conceptos para pensar la realidad, a veces los confundimos a los dos. Tomamos erróneamente a los hechos conceptuales por hechos sobre la realidad. Y, en el caso de ciertos conceptos, los que se hallan más cargados de significación emocional o moral, podemos despistarnos seriamente. De estos conceptos cargados, ese de nuestra propia identidad es, tal vez, el que más induce a confusión.

Hasta el uso de la palabra «yo» puede despistarnos. Consideremos el hecho de que, dentro de unos cuantos años, yo estaré muerto. Este hecho puede parecer deprimente. Pero la realidad es solamente esta: tras un cierto tiempo, ninguno de los pensamientos y ninguna de las experiencias que ocurren estarán directa y causalmente relacionados con este cerebro, ni estarán conectados de ciertos modos con estas experiencias actuales. Eso es todo lo que este hecho conlleva. Y, en esta redescripción, mi muerte parece desaparecer.

BIBLIOGRAFÍA

ADAMS (1), R. M., «Must God Create the Best?», *Philosophical Review* 81, n.º 3, julio 1972.

ADAMS (2), R. M., «Motive Utilitarianism», *The Journal of Philosophy*, 12 agosto 1976.

ADAMS (3), R. M., «Existence, Self-Interest and the Problem of Evil», *Nous*, 13, 1979.

ANSCHUTZ, R. P., *The Philosophy of J. S. Mill*, Oxford, Clarendon Press, 1953.

ANSCOMBE (1), G. E. M., *Intention*, Ithaca, N. Y., Cornell University Press, 1957. (Versión castellana de A. I. Stellino: *Intención*, con introducción de Jesús Mosterín, en Paidós/ICE-UAB, Barcelona, 1991.)

ANSCOMBE (2), G. E. M., «Who is Wronged?», *The Oxford Review*, 1967.

AYER (1), A. J., *The Concept of a Person and Other Essays*, Londres, Macmillan, 1964. (Versión castellana: *El concepto de persona*, en Seix Barrall, Barcelona, 1966.)

AYER (2), A. J., *Philosophical Essays*, Londres, Macmillan, 1965. (Versión castellana: *Ensayos filosóficos*, en Ariel, Barcelona, 1970.)

BAIER (1), K, *The Moral Point of View*, Ithaca, N. Y., Cornell University Press, 1958.

BAIER (2), K., «Ethical egoism and Interpersonal Compatibility», *Philosophical Studies*, 24, 1973.

BARRY (1), B., *Political Argument*, Londres, Routledge and Kegan Paul, 1965.

BARRY (2), B., «Justice Between Generations», en P. M. S. Hacker y J. Raz (eds.), *Law, Morality and Society: Essays in Honour of H. L. A. Hart*, Oxford, Clarendon Press, 1977.

BARRY (3), B., «Rawls on Average and Total Utility: A Comment», *Philosophical Studies*, 31, 1977.

BARRY (4), B., *Sociologists, Economists, and Democracy*, Londres, 1970.

BAYLES, M. D. (ed.), *Ethics and Population*, Cambridge, Mass., Schenkman, 1976.

BENDITT, T., «Happiness», *Philosophical Studies*, 25, 1974.

BENNETT, J., *Kant's Analytic*, Cambridge University Press, 1966. (Versión castellana de A. Montesinos: *La crítica de la razón pura de Kant. v. 1, La analítica*, en Alianza, Madrid, 1979.)

BENTHAM, J., *An Introduction to the Principles of Morals and Legislation*, publicado por vez primera en 1789, y reeditado en *Utilitarianism*, M. Warnock, ed., Londres, Collins, 1962.

BLANSHARD, B., «Sidgwick The Man», *The Monist*, 1974.

BOGEN, J., «Identity and Origin», *Analysis*, 26, abril 1966.

BRAMS, S. J., «Newcomb's Problem and Prisioners' Dilemma», *Journal of Conflict Resolution*, 19, n.° 4, diciembre 1975.

BRANDT (1), R. B., ed., *Value and Obligation*, Nueva York, Harcourt, Brace and World, 1961.

BRANDT (2), R. B., *A Theory of the Good and the Right*, Oxford, Clarendon Press, 1979.

BRAYBROOKE, D., «The Insoluble Problem of the Social Contract», *Dialogue*, marzo 1976.

BRENNAN (1), A. A., «Personal Identity and Personal Survival», *Analysis*, 42, enero 1982.

BRENNAN (2), A. A., «Survival», *Synthese*, 1984.

BROAD (1), C. D., «On the Function of False Hypotheses in Ethics», *Ethics*, 1915.

BROAD (2), C. D., *The Mind and its Place in Nature*, Londres, Routledge and Kegan paul, 1949.

BROAD (3), C. D., *Five Types of Ethical Theory*, Littlefield, Adams, 1959.

BROOME (1), J., «Indefiniteness in Identity», *Analysis*, en o antes de 1984.

BROOME (2), J., «Rational Choice and Value in Economics», *Oxford Economic Papers*, 30, noviembre 1978.

BUCHANAN, A., «Revolutionary Motivation and Rationality», *Philosophy and Public Affairs*, 9, n.° 1, otoño 1979.

BUTLER, J., *The Analogy of Religion*, Primer Apéndice, 1736, reeditado en Perry (1).

CHISHOLM, R., «Reply to Strawson's Comments», en H. E. Kiefer y Milton K. Munitz, eds., *Language, Belief, and Metaphysics*, Albany, N. Y., State University of New York Press, 1970.

COLLINS, S., *Selfless Persons*, Cambridge University Press, 1982.

DANIELS, N., «Moral Theory and the Plasticity of Persons», *The Monist*, 62, n.° 3, julio 1979.

DENNETT, D. C., *Brainstorms*, Bradford Books, 1978.

DESCARTES, R., *Meditations*, traducidas por E. S. Haldane y G. R. T. Ross, Cambridge University Press, 1969. (Versión castellana, con introducción y notas de Vidal Peña: *Meditaciones metafísicas con objeciones y respuestas*, Alfaguara, Madrid, 1977.)

DUMMETT, M. A. E., «Wang's Paradox», *Synthese*, 30, 1975.

DWORKIN (1), R., «What is Equality? Part 1: Equality or Welfare», *Philosophy and Public Affairs*, 10, n.° 3, verano 1981.

DWORKIN (2), R., «What is Equality? Part 2: Equality of Resources», *Philosophy and Public Affairs*, 10, n.° 4, otoño 1981.

EDGELY, R., *Reason in Theory and Practice*, Londres, Hutchinson, 1969.

EDIDIN, A., «Temporal Neutrality and Past Pains», *The Southern Journal of Philosophy*, 20, n.° 4, invierno 1982.

EDWARDS, R. B., *Pleasures and Pains*, Ithaca, N. Y., Cornell University Press, 1979.

ELLIOT, R., «How to Travel Faster Than Light?», *Analysis*, 41, enero 1981.

EVANS (1), G., «Can There Be Vague Objects?», *Analysis*, 38, 1978.

EVANS (2), G., *The Varieties of Reference*, Oxford, Clarendon Press, 1982.

EWING, A. C., «Suppose Everybody Acted Like Me», *Philosophy,* 28, enero 1953.

FEINBERG, J., «The Rights of Animals and Future Generations», en William K. Blackstone, ed., *Philosophy and Environmental Crisis*, Athens, Ga, University of Georgia Press, 1974.

FINDLAY, J., *Values and Intentions*, Londres, George Allen and Unwin, 1961.

FISHKIN (1), J. S., *The Limits of Obligation*, New Haven, Ct., Yale University Press, 1982.

FISHKIN (2), J. S., «Justice Between Generations: the Dilemma of Future Interests», *Social Justice,* 4, 1982.

FOOT, P., «Morality as a System of Hypothetical Imperatives», *The Philosophical Review,* 81, 1972.

FORBES (1), G., «Origin and Identity», *Philosophical Studies,* 37, 1980.

FORBES (2), G., «Thisness and Vagueness», *Synthese,* 54, 1983.

GALE (1), R. M., *The Language of Time*, Londres, Macmillan, 1968.

GALE (2), R. M., ed., *The Philosophy of Time*, Londres, Macmillan, 1968.

GAUTHIER (1), D., «Morality and Advantage», *Philosophical Review*, 1967.

GAUTHIER (2), D., *Practical Reasoning*, Oxford, Clarendon Press, 1962.

GAUTHIER (3), D., «Reason and Maximization», *Canadian Journal of Philosophy*, marzo 1975.

GAUTHIER (4), D., *Morals by Agreement* (título provisional), Oxford, Clarendon Press, de próxima publicación. (Versión castellana de Alcira Bixio: *La moral por acuerdo*, en Gedisa, Barcelona, 1994.)

GERT, B., *The Moral Rules*, Nueva York, Harper and Row, 1973.

GLOVER (1), J. C. B., *Causing Death and Saving Lives*, Harmondsworth, Penguin Books, 1977.

GLOVER (2), J. C. B., «It Makes No Difference Whether Or Not I Do It», *Proceedings of the Aristotelian Society, Suplem.* Vol. 49, 1975.

GLOVER (3), J. C. B., *What Sort of People Should There Be?*, Harmondsworth, Penguin Books, 1984.

GODWIN, W., *Political Justice*, 1793, y Oxford University Press, 1971. (Versión castellana de J. Prince: *Investigación acerca de la justicia política*, en Júcar, Madrid, 1985.)

GOODIN, R. E., «Discounting Discounting», *Journal of Public Policy,* 2, Pt. 1 febrero 1982.

GOSLING, J. C. B., *Pleasure and Desire*, Oxford, Clarendon Press, 1969.

GRICE, G. R., *The Ground of Moral Judgement*, Cambridge University Press, 1967.

GRIFFIN (1), J. P., «Are There Incommensurable Values?», *Philosophy and Public Affairs,* 7, n.º 1, otoño 1977.

GRIFFIN (2), J. P., «Modern Utilitarianism», *Revue Internationale de Philosophie*, n.º 141, 1982.

GRIFFIN (3), J. P., «A Substantive Theory of Rights», artículo inédito.

GRIFFIN (4), J. P., «Is Unhappiness Morally More Important Than Happiness?», *Philosophical Quarterly,* 29, 1979.

GRIM, P., «What Won't Escape Sorites Arguments», *Analysis,* 42, 1982, p. 38.

GUTTENPLAN, S., ed., *Mind and Language*, Oxford, Clarendon Press, 1975.

HAIGHT, G. S., *George Eliot: A Biography*, Oxford University Press, 1978.

HAKSAR, V., *Equality, Liberty and Perfectionism*, Oxford, Clarendon Press, 1979.

HARDIN, G., «The Tragedy of the Commons», *Science,* 162, 13, diciembre 1968.

HARE (1), R. M., *Moral Thinking*, Oxford, Clarendon Press, 1981.

HARE (2), R. M., «Ethical theory and Utilitarianism», en H. D. Lewis, ed., *Contemporary British Philosophy*, Londres, George Allen and unwin, 1976, reeditado en Sen y Williams.

HARE (3), R. M., «Pain and Evil», *Proceedings of the Aristotelian Society*, *Suplem.* Vol. 38, 1964.

HARE (4), R. M., *Freedom and Reason*, Oxford University Press, 1963.

HARE (5), R. M., «Abortion and the Golden Rule», *Philosophy and Public Affairs*, 4, n.º 3, primavera 1975.

HARE (6), R. M., *Essays on Philosophical Method*, Londres, Macmillan, 1971.

HARE (7), R. M., *Practical Inferences*, Londres, Macmillan, 1971.

HARE (8), R. M., *Essays on the Moral Concepts*, Londres, Macmillan, 1972.

HARE (9), R. M., *Applications of Moral Philosophy*, Londres, Macmillan, 1972.

HARMAN, G., *The Nature of Morality*, Oxford University Press, 1977. (Versión castellana de Cecilia Hidalgo revisada por Eduardo Rabossi: *La naturaleza de la moralidad: una introducción a la ética*, en México, Universidad Nacional Autónoma, 1983.)

HOFSTADTER y DENNETT: D. R. Hofstadter y D. C. Dennett, eds., *The Mind's I*, Brighton, Harvester Press, 1981. (Versión castellana de Lucrecia de Sáenz: *El ojo de la mente: fantasías y reflexiones sobre el yo y el alma*, en Editorial Sudamericana, Buenos Aires, 1983.)

HOLLIS (1), M., «Rational Man and Social Science», en R. Harrison, ed., *Rational Action: Studies in Philosophy and Social Science*, Cambridge University Press, 1979.

HOLLIS (2), M., *Models of Man*, Cambridge University Press, 1977.

HUME (1), D., *A Treatise of Human Nature*, Oxford, Clarendon Press, 1978. (Versión castellana en edición preparada por Félix Duque: *Tratado de la naturaleza humana: autobiografía*, en Editora Nacional, Madrid, 1977.)

HUME (2), D., «The Sceptic», en los *Essays* de Hume, Oxford University Press, 1963.

HUME (3), D., *An Enquiry Concerning the Principles of Morals*, Oxford, Clarendon Press, 1975. (Versión castellana de Carlos Mellizo, con prólogo y notas: *Investigación sobre los principios de la moral*, en Alianza, Madrid, 1993.)

HURKA (1), T. M., «Average Utilitarianisms», *Analysis*, 42, n.º 2, marzo 1982.

HURKA (2), T. M., «More Average Utilitarianisms», *Analysis*, 42, junio 1982.

HURKA (3), T. M., «Value and Population Size», *Ethics*, 93, n.º 3, abril 1983.

KANT, I., *Critique of Pure Reason*, traducción de N. Kemp Smith, Londres, Macmillan, 1964. (Versión castellana, con prólogo notas e índices, de Pedro Rivas: *Crítica de la Razón Pura*, en Alfaguara, Madrid, 1978.).

KAVKA (1), G., «Rawls on Average and Total Utility», *Philosophical Studies*, 27, abril 1975.

KAVKA (2), G., «Some Paradoxes of Deterrence», *The Journal of Philosophy*, 75, n.º 6, junio 1978.

KAVKA (3), G., «Deterrence, Utility, and Rational Choice», *Theory and Decision*, 12, 1980.

KAVKA (4), G., «The Paradox of Future Individuals», *Philosophy and Public Affairs*, 11, n.º 2, primavera 1982.

KRIPKE, S. A., «Naming and Necessity», en G. Harman y D. Davidson, eds., *Semantics of Natural Language*, Dordrecht, Reidel, 1972.

LESLIE, J., *Value and Existence*, Oxford, Basil Blackwell, 1979.

LEWIS (1), C. I., *An Analysis of Knowledge and Valuation*, La Salle, Ill., Open Court, 1946.

LEWIS (2), D. K., *Convention: A Philosophical Study*, Cambridge, Mass., Harvard University Press, 1969.

LEWIS (3), D. K., «Survival and Identity», en Rorty.

LEWIS (4), H. D., *The Elusive Mind*, Londres, George Allen and Unwin, 1969.

LEWIS (5), H. D., *The Self and Immortality*, Nueva York, Seabury Press, 1973.

LEWIS (6), H. D., *The Elusive Self*, Londres, Macmillan, 1982.

LICHTENBERG, G. C., *Schriften und Briefe*, Sudelbucher II, Carl Hanser Verlag, 1971.

LOCKE, J., *Essay Concerning Human Understanding*, reeditado parcialmente en Perry (1). (Versión castellana de María Esmeralda García, en edición preparada por Sergio Rábade para Editora Nacional, Madrid, 1980.)

LYONS (1), D., *Forms and Limits of Utilitarianism*, Oxford, Clarendon Press, 1965.

LYONS (2), D., «Human Rights and the General Welfare», *Philosophy and Public Affairs*, 6, n.º 2, invierno 1979.
MCDERMOTT, M., «Utility and Population», *Philosophical Studies*, 42, 1982.
MACKAYE, J., *The Economy of Happiness*, Boston, Little, Brown, 1906.
MACKIE (1), J. L., «Sidgwick's Pessimism», *Philosophical Quarterly*, 1976.
MACKIE (2), J. L., *Ethics*, Harmondsworth, Penguin Books, 1977. (Versión castellana de Tomás Fernández Azúa: *Ética: la invención de lo bueno y lo malo*, en Gedisa, Barcelona, 2000.)
MACKIE (3), J. L., «Rules and Norms», en vol. 1 de *Selected Papers*, Oxford, Clarendon Press, de próxima publicación.
MACKIE (4), J. L., *Problems from Locke*, Oxford, Clarendon Press, 1976.
MACKIE (5), J. L. «The Transcendental 'I'», en Van Straaten.
MACLEAN y BROWN: D. Maclean y P. G. Brown, eds., *Energy and the Future*, Totowa, N. J. Rowman and Littlefield, 1983.
MCMAHAN (1), J. A., «Problems of Population Theory», *Ethics*, 92, n.º 1, octubre 1981.
MCMAHAN (2), J. A., «Nuclear Deterrence and Future Generations», en S. Lee y A. Cohen, eds., *Nuclear Weapons and the Future of Humanity*, Paterson, N. J., Littlefield Adams, de próxima publicación.
MADELL, G., *The Identity of the Self*, Edinburgh University Press, 1981.
MARGLIN, S., «The Social Rate Discount and the Optimal Rate of Investment», *Quarterly Journal of Economics*, 1963.
MARTIN y DEUTSCHER, «Remembering», *Philosophical Review*, 1966.
MEEHL, P., «The Paradox of the Throw-Away Vote», *American Political Science Review*, 71, 1977.
MILLER y SARTORIUS: F. Miller y R. Sartorius, «Population Policy and Public Goods», *Philosophy and Public Affairs*, 8, n.º 2, invierno 1979.
MILLER y WILLIAM: H. B. Miller y W. H. Williams, eds., *The Limits of Utilitarianism*, Minneapolis, University of Minnesota Press, 1982.

MONTEFIORE, A., ed., *Philosophy and Personal Relations*, Londres, Routledge and Kegan Paul, 1973.
MOORE, G. E., *Principia Ethica*, Cambridge University Press, 1903. (Versión castellana, con introducción, de Tomas Baldwin, en Universidad Nacional Autónoma de México, 1997, edición revisada y ampliada con el prefacio a la 2.ª edición.)
NABOKOV, V., *Glory*, Londres, Weidenfeld y Nicholson, 1971.
NAGEL (1), T., *The Possibility of Altruism*, Oxford University Press, 1970.
NAGEL (2), T., «War and Massacre», *Philosophy and Public Affairs*», 1, n.º 2, invierno 1972.
NAGEL (3), T., «The Limits of Objectivity», en *The Tanner Lectures on Human values*, university of Utah Press, 1980.
NAGEL (4), T., *Mortal Questions*, Cambridge University Press, 1979. (Versión española: *La muerte en cuestión*, en Fondo de Cultura Económica, México, 1981.)
NAGEL (5), T., «Brain Bisection and the Unity of Consciousness», *Synthese*, 22, 1971, reeditado en Nagel (4) y en Perry (1).
NARVESON (1), J., «Utilitarianism and New Generations», *Mind*, 76, enero 1967.
NARVESON (2), J., «Moral problems of Population», *The Monist*, 57, n.º 1, enero 1973.
NEWMAN, Cardenal Newman, *Certain Difficulties Felt by Anglicans in Catholic Teaching*, Londres, 1885.
NIETZSCHE, F., *The Portable Nietzsche*, editado y traducido por W. Kaufman, Nueva York, Viking Press, 1963.
NORMAN, R., *Reasons for Actions*, Oxford, Basil Blackwell, 1971.
NOZICK (1), R., «On Austrian Methodology».
NOZICK (2), R., *Anarchy, State, and Utopia*, Oxford, Basil Blackwell, 1974. (Versión española de Rolando Tamayo: *Anarquía, estado y utopía*, en Fondo de Cultura Económica, 1988.)
NOZICK (3), R., *Philosophical Explanations*, Cambridge, Mass., Harvard University Press, 1981.
OLSON, M., *The Logic of Collective Action*, Cambridge, Mass., Harvard University Press, 1965. (Versión castellana de Ricardo Calvet Pérez: *La lógica de la acción colectiva: bienes públicos y la teoría de los grupos*, en Limusa, México, 1992.)

PARFIT (1), D., «Personal Identity», *Philosophical Review,* 80, n.º 1, enero 1971, reeditado en Perry (1). (Versión castellana de Álvaro Rodríguez Tirado: *Identidad personal*, en UNAM/ Instituto de Investigaciones Filosóficas, Cuadernos de Crítica, México, 1983.)
PARFIT (2), D., «On The Importance of Self-Identity», *The Journal of Philosophy,* 68, n.º 20, 21 octubre 1971.
PARFIT (3), D., «Later Selves and Moral Principles», en Montefiore.
PARFIT (4), D., «Innumerate Ethics», *Philosophy and Public Affairs,* 7, n.º 4, verano 1978.
PARFIT (5), D., «Prudence, Morality, and the Prisioner's Dilemma», *Proceedings of the British Academy,* 65, 1979. (Versión castellana de Gilberto Gutiérrez: *Prudencia, moralidad y el Dilema del Prisionero,* en *Excerpta Philosophica,* 2, Facultad de Filosofía de la Universidad Complutense de Madrid.)
PARFIT (6), D., «Personal Identity and Rationality», *Synthese,* 53, 1982.
PARFIT (7), D., «Future Generations: Further Problems», *Philosophy and Public Affairs,* 11, n.º 2, primavera 1982.
PARTRIDGE, E., ed., *Responsabilities to Future Generations,* Buffalo, N. Y., Prometheus Books, 1981.
PEACOCKE (1), C., «Are Vague Predicates Incoherent?», *Synthese,* 46, 1981.
PEACOCKE (2), C., *Sense and Content*, Oxford, Clarendon Press, 1983.
PEARS (1), D. F., «Time, Truth, and inference», en A. Flew, ed., *Essays in Conceptual Analysis*, Londres, Macmillan, 1966.
PEARS (2), D. F., «Critical Study of *Individuals*, by P. F. Strawson», *Philosophical Quarterly,* 11, 1961.
PENELHUM, T., «The Importance of Self-identity», *The Journal of Philosophy,* 68, n.º 20, 21 octubre 1971.
PERRY (1), J., ed., *Personal Identity*, Berkeley, University of California Press, 1975.
PERRY (2), J., *A Dialogue on Personal identity and Immortality*, Indianapolis, Hackett, 1978. (Versión española de Ariel Campiran: *Diálogo sobre la identidad personal y la inmortalidad*, en UNAM/

Instituto de Investigaciones Filosóficas, Cuadernos de crítica, México, 1984.)
PERRY (3), R. B., *General Theory of Value*, Cambridge, Mass., Harvard University Press, 1950.
PLATÓN (1), *Protágoras*. (Traducción castellana en Gredos, Madrid, 1982: *Diálogos*, vol. 1, introducción general de Emilio Lledó.)
PLATÓN (2), *Filebo*. (Edición citada, vol. 6.)
PROUST (1), M., *The Sweet Cheat Gone*, 1, traducción de C. K. Scott-Moncrieff, Londres, Chatto and Windus, 1949.
PROUST (2), M., *Within a Budding Grove*, traducción de C. K. Scott-Moncrieff, Londres, Chatto and Windus, 1949.
QUINE (1), W. V., reseñando a Milton K. Munitz, ed., *Identity and Individuation*, en *The Journal of Philosophy*, 1972.
QUINE (2), W. V., «What Price Bivalence?», *The Journal of Philosophy*, 1981.
QUINTON, A., «The Soul», *The Journal of Philosophy*, 59, n.º 15, julio 1962, reeditado en Perry (1).
RAHULA, W., *What the Buddha Taught*, Nueva York, Grove Press, 1974.
RAILTON, P., «Alienation and the Demands of Morality», *Philosophy and Public Affairs*, de próxima publicación.
RAPOPORT, A., *Fights, Games, and Debates*, Ann Arbor, University of Michigan Press, 1960.
RASHDALL, H., *The Theory of Good and Evil*, Oxford University Press, 1907.
RAVERAT, G., *Period Piece*, Londres, Faber and Faber, 1952.
RAWLS, J., *A Theory of Justice*, Cambridge, Mass., Harvard University Press, 1971. (Versión castellana de María Dolores González: *Teoría de la justicia*, en Fondo de Cultura Económica, México, 1979.)
RAY, C., «Can We Travel Faster Than Light?», *Analysis,* 42, enero 1982.
RAZ, J., *Practical Reason and Norms*, Londres, Hutchinson, 1975. (Versión castellana de Juan Ruiz Manero: *Razón práctica y normas*, en Centro de Estudios Constitucionales, Madrid, 1991.)
REGAN, D., *Utilitarianism and Co-operation*, Oxford, Clarendon Press, 1980.

REID, T., *Essays on the Intellectual Powers of Man*, publicado por vez primera en 1785, «Of Memory», cap. 4, reeditado en Perry (1).
RESCHER, N., *Unselfishness*, University of Pittsburg Press, 1975.
RISCHARDS, D. A. J., *A Theory of Reasons for Action*, Oxford, Clarendon Press, 1971.
RIKER y ORDESHOOK: W. Riker y P. Ordeshook, «A Theory of the Calculus of Voting», *American Political Science Review*, 62, marzo 1978.
ROBERTSON, Sir D., *Lectures on Economic Principles*, Londres, Collins, 1969.
RORTY, A., ed., *The Identities of Persons*, Berkeley, University of California Press, 1976.
ROSS (1), W. D., *The Right and the Good*, Oxford, The Clarendon Press, 1930. (Versión castellana de Leonardo Rodríguez: *Lo correcto y lo bueno*, en Sígueme, Salamanca, 1994.)
ROSS (2), W. D., *The Foundations of Ethics*, Oxford, The Clarendon Press, 1939.
RUSSELL, B., «On the Nature of Acquaintance», reeditado en R. C. Marsh, ed., *Logic and Knowledge*, Londres, Allen and Unwin, 1956. [Versión castellana de Javier Muguerza: *Ensayos sobre lógica y conocimiento (1901-1950)*, comp. por R. C. Marsh, en Taurus, Madrid, 1966.]
SAINSBURY, R. M., «In Defence of Degrees of Truth», artículo inédito.
SALMON, N. U., *Reference and Essence*, Oxford, Basic Blackwell, 1982.
SAMUELSON, P. A., *Economics*, New York, McGraw-Hill, 1970. (Versión castellana de M. Gala, D. Azqueta y L. Toharia: *Economía*, en McGraw-Hill, Madrid, 1983, 11.ª ed.)
SCANLON (1), T. M., «Preference and Urgency», *The Journal of Philosophy*, 72, n.º 19, 6 noviembre 1975.
SCANLON (2), T. M., «Rights, Goals, and Fairness», en S. Hampshire, ed., *Public and Private Morality*, Cambridge University Press, 1978.
SCHEFFLER (1), S., «Moral independence and the Original Position», *Philosophical Studies*, 35, 1979.
SCHEFFLER (2), S., *The Rejection of Consequentialism*, Oxford, Clarendon Press, 1982.

SCHEFFLER (3), S., «Ethics, Personal Identity, and the Ideals of the Person», *Canadian Journal of Philosophy*, 12, n.º 2, junio 1982.
SCHELL, J., *The Fate of the Earth*, Nueva York, Avon Books, 1982.
SCHELLING (1), T., *The Strategy of Conflict*, Cambridge, Mass, Harvard University Press, 1960. (Versión castellana: *La estrategia del conflicto*, en Tecnos, Madrid, 1964.)
SCHELLING (2), T., «Hockey Helmets, Concealed Weapons, and Daylight Saving», *The Journal of Conflict Resolution*, septiembre 1973.
SCHNEEWIND, J. B., *Sidgwick's Ethics and Victorian Moral Philosophy*, Oxford University Press, 1977.
SCHUELER, G. F., «Nagel on the Rationality of Prudence», *Philosophical studies*, 29, 1976.
SEN (1), A. K., *Collective Choice and Social Welfare*, San Francisco, Holden Day, 1970. (Versión castellana de F. Elías Castillo: *Elección colectiva y bienestar social*, en Alianza, Madrid, 1976.)
SEN (2), A. K., *Behaviour and the Concept of Preference*, London School of Economics, 1973.
SEN (3), A. K., «Isolation, Assurance, and the Social Rate of Discount», *Quarterly Journal of Economics*, 81, 1967.
SEN (4), A. K., «Rational Fools: A Critique of the Behavioral Foundations of Economic Theory», *Philosophy and Public Affairs*, 6, n.º 4, verano 1977.
SEN (5), A. K., «Utilitarianism and Welfarism», *The Journal of Philosophy*, 76, n.º 9, septiembre 1979.
SEN (6), A. K., *On Economic Inequality*, Oxford, Clarendon Press, 1973. (Versión castellana: *La desigualdad económica*, en Fondo de Cultura Económica, México, 2001, edición ampliada.)
SEN (7), A. K., «Choice, Orderings, and Morality», en S. Korner, ed., *Practical Reason*, Oxford University Press, 1974.
SEN y RUNCIMAN: A. K. Sen y W. G. Runciman, «Games, Justice and the General Will», *Mind*, 74, octubre 1965.
SEN y WILLIAMS: A. K. Sen y B. Williams, *Utilitarianism and Beyond*, Cambridge University Press, 1982.
SHOEMAKER (1), S., *Self-Knowledge and Self-Identity*, Ithaca, N. Y., Cornell University Press, 1963.

SHOEMAKER (2), S., «Persons and Their Pasts». *American Philosophical Quarterly*, 7, 1970. (Versión castellana de J. Lascuráin y E. Villanueva: *Las personas y su pasado* en Instituto de Investigaciones Filosóficas, México, 1981.)

SHOEMAKER (3), S., «Wiggins on Identity», *Philosophical Review*, 79, 1970. (Versión castellana de M.ª Isabel Martínez y E. Villanueva: *Wiggins y la identidad y La persistencia de las personas*, en Instituto de Investigaciones Filosóficas, México, 1986.)

SHORTER, J. M., «More About Bodily Continuity and Personal Identity», *Analysis*, 22, 1961-2.

SIDGWICK (1), H., *The Methods of Ethics*, Londres, Macmillan, 1907.

SIDGWICK (2), A. S. y E. M. S. Sigdwick, *Henry Sidgwick, a Memoir*, Londres, Macmillan, 1906.

SIKORA y BARRY: R. I. Sikora y B. Barry, eds., *Obligations to Future Generations*, Philadelphia, Temple University Press, 1978.

SINGER (1), M., «The Many Methods of Sidgwick's Ethics», *The Monist*, 58, 1974.

SINGER (2), P., *The Expanding Circle*, Nueva York, Farrar, Straus y Giroux, 1981.

SMART (1), B., «Diachronus and Synchronus Selves», *Canadian Journal of Philosophy*, 6, n.° 1, marzo 1976.

SMART (2), J. J. C., ed., *Problems of Space and Time*, Nueva York, Macmillan, 1976.

SMART y WILLIAMS: J. J. C. Smart y B. Williams, *Utilitarianism: For and Against*, Cambridge University Press, 1973. (Versión castellana: *Utilitarismo, pro y contra*, en Tecnos, Madrid, 1981.)

SOBEL (1), J. H., «Interaction Problems for Utility Maximizers», *Canadian Journal of Philosophy*, 4, n.° 4, junio 1975.

SOBEL (2), J. H., «The Need for Coercion», en J. Pennock y H. Chapman, eds., *Coercion*, Chicago, Nomos, 1972.

SOLZHENITSYN, A., *The First Circle*, Nueva York, Bantham Books, 1969.

SPERRY, R. W., en J. Eccles, ed., *Brain and Conscious Experience*, Berlín, Springer Verlag, 1966.

STRANG, C., «What if Everyone Did That?», en Thomson y Dworkin.

STRAWSON (1), P. F., *Individuals*, Londres, Methuen, 1959. (Versión castellana de A. García Suárez y L. M. Valdés Villanueva: *Individuos: ensayo de metafísica descriptiva*, en Taurus, Madrid, 1987.)

STRAWSON (2), P. F., *The Bounds of Sense*, Londres, Methuen, 1966. (Versión castellana de Carlos Thiebaut: *Los límites del sentido: ensayo sobre la «Crítica de la razón pura» de Kant*, en Ediciones de la Revista de Occidente, Madrid, 1975.)

STROTZ: «Inconsistency and Myopia in Dynamic Utility Maximization», *Review of Economic Studies*, 1955-6.

SUMNER (1), L. W., «The Good and the Right», *Canadian Journal of Philosophy*, Suplem. Vol. sobre John Stuart Mill y el Utilitarismo, 1979.

SUMNER (2), L. W., *Abortion and Moral Theory*, Princeton University Press, 1981.

SWINBURNE, R. G., «Personal Identity», *Proceedings of the Aristotelian Society*, 74, 1973-4.

TEMKIN, L., «Inequality», Tesis Doctoral, Universidad de Princeton, 1982.

THOMSON y DWORKIN: J. J. Thomson y G. Dworkin, eds., *Ethics*, Nueva York, Harper and Row, 1968.

TOOLEY, M., *Abortion and Infanticide*, Oxford, Clarendon Press, 1983.

TREBLICOT, J., «Aprudentialism», *American Philosophical Quarterly*, 11, n.° 3, julio 1974.

ULLMAN-MARGALIT, E., *The Emergence of Norms*, Oxford, Clarendon Press, 1977.

UNGER (1), P., «Why There Are No People», *Midwest Studies in Philosophy*, 4, 1979.

UNGER (2), P., «I Do Not Exist», en G. F. MacDonald, ed., *Perception and Identity*, Ithaca, N. Y., Cornell University Press, 1979.

VAN STRAATEN, Z., *Philosophical Subjects, Essays presented to P. F. Strawson*, Oxford, Clarendon Press, 1980.

WACHSBERG, M., *Personal Identity, the Nature of Persons, and Ethical Theory*, Tesis Doctoral, Universidad de Princeton, 1983.

WARNOCK, G. J., *The Object of Morality*, Londres, Methuen, 1971.

WATKIN, J., «Imperfect Rationality», en R. Borger y F. Cioffi, eds., *Explanation in the Behavioural Sciences*, Cambridge University Press, 1970.
WIGGINS (1), D., *Identity and Spatio-Temporal Continuity*, Oxford, Basil Blackwell, 1967.
WIGGINS (2) D., «Essentialism, Continuity, and Identity», *Synthese,* 23, 1974.
WIGGINS (3), D., *Sameness and Substance*, Oxford, Basil Blackwell, 1980.
WIGGINS (4), D., «Locke, Butler and the Stream of Consciousness», en Rorty. (Versión castellana: *Locke, Butler y la corriente de conciencia: los hombres como una clase natural*, en Instituto de Investigaciones Filosóficas, México, 1986.)
WIGGINS (5), D., «Identity, Designation, Essentialism and Physicalism», *Philosophia,* 5, n.ºs 1-2, enero-abril 1975.
WIGGINS (6), D., «The Concern to Survive», *Midwest Studies in Philosophy,* 4, 1979.
WILLIAMS (1), B., «A Critique of Utilitarianism», en Smart y Williams.

WILLIAMS (2), B., *Problems of the Self*, Cambridge University Press, 1973. (Versión castellana de José M. G. Holguera: *Problemas del yo*, en Universidad Nacional Autónoma, México, 1986.)
WILLIAMS (3), B., «Persons, Character, and Morality», en Rorty.
WILLIAMS (4), B., *Descartes*, Harmondsworth, Penguin Books, 1978. (Versión castellana de Laura Benítez: *Descartes: el proyecto de la investigación pura*, en Universidad Nacional Autónoma/Instituto de Investigaciones Filosóficas, México, 1995.)
WILLIAMS (5), B., «Internal and External reasons», en R. Harrison, ed., *Rational Action: Studies in Philosophy and Social Science*, Cambridge University Press, 1979.
WILLIAMS (6), B., *Moral Luck*, Cambridge University Press, 1981. (Versión castellana de Susana Martín: *La fortuna moral: ensayos filosóficos 1973-1980*, en Universidad Nacional Autónoma, México, 1993.)
WILLIAMS (7), B., «The Point of View of the Universe: Sidgwick and the Ambitions of Ethics», *The Cambridge Review* 7, mayo 1982.

WILLIAMS (8), B., «The Self and the Future», *Philosophical Review,* 79, n.° 2, abril 1970, reeditado en Williams (2).
WILLIAMS (9), B., «Bodily Continuity and Personal Identity», *Analysis,* 20, n.° 5, reeditado en Williams (2).
WOODS, M., «Reference and Self-Identification», *The Journal of Philosophy*, 1968.

ÍNDICE DE NOMBRES

* Los números en cursiva remiten a nota en las páginas mencionadas.

 905

W